REVIEW of EDUCATIONAL RESEARCH

Volume III FEBRUARY, 1933 Number 1

EDUCATIONAL TESTS AND THEIR USES

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
A Department of the
NATIONAL EDUCATION ASSOCIATION
1201 SIXTEENTH STREET NORTHWEST
WASHINGTON, D. C.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

THIS Association is composed of persons who are engaged in technical research in education, including directors of research in school systems, instructors in educational institutions, and research workers connected with private educational agencies. The Association became a Department of the National Education Association in July, 1930.

Officers of the Association for 1932-33

President
WILLIAM S. GRAY
University of Chicago

Vice-President
ALFRED D. SIMPSON
New York State Department of Education

Secretary-Treasurer
WILLIAM G. CARR
Research Division
National Education Association

Executive Committee

The president, vice-president, and secretary, *ex officio*, and the following past presidents:

W. W. CHARTERS
Ohio State University

J. L. STENQUIST
Baltimore Public Schools

EDITORIAL BOARD, 1932-33

FRANK N. FREEMAN, *Chairman*, University of Chicago
J. CAYCE MORRISON, State Department of Education, Albany, New York
HARRY J. BAKER, Detroit Public Schools
President and Secretary, *ex officio*

Active Membership—Persons eligible to membership in the Association must be recommended to the Executive Committee by a member who is in good standing. Upon approval of the recommendation, the person recommended will be invited by the Executive Committee to become a member of the Association. The Executive Committee has defined the qualifications for membership as follows:

"Membership in the Association is restricted to persons of good ability and sound training who are working in the field of educational research, and who can present satisfactory evidence in the form of published or unpublished studies which show ability to arrange, to organize, and to conduct research investigations and experiments. In addition, evidence of an abiding interest in the field of educational research is essential."

Membership in the National Education Association is a prerequisite to active membership in the American Educational Research Association. Any form of N. E. A. membership—annual, five-dollar, or life—satisfies the preliminary requirement.

Active members of the Association pay dues of $5.00 per year. Of this amount, $4.00 is for subscription to THE REVIEW. See back inside cover page of this issue. THE REVIEW is published in February, April, June, October, and December each year.

Entered as second-class matter April 10, 1931, at the post office at Washington, D. C., under the Act of August 24, 1912.

REVIEW OF EDUCATIONAL RESEARCH

Official Publication of the American Educational Research Association, a department of the National Education Association.

The contents of the Review are listed in the EDUCATION INDEX

Volume III **February, 1933** **Number 1**

EDUCATIONAL TESTS AND THEIR USES

(Literature reviewed to December 1, 1932)

Prepared by the Committee on Educational Tests and Their Uses: Ben D. Wood, W. J. Osburn, G. M. Ruch, M. R. Trabue, Grace A. Kramer, and John L. Stenquist, Chairman; with the assistance of E. F. Lindquist and H. R. Anderson.

TABLE OF CONTENTS

FOREWORD

THE field of tests as a subject of review is less clearly defined than are the fields which deal with content as distinguished from method. One may approach the subject in a variety of ways. One may, for example, make a descriptive summary of the various types of tests. Again, one may discuss the technical problems and review the technical studies dealing with the construction and use of tests. Still another method of dealing with the subject is to discuss the practical ways in which tests may be used in dealing with problems of administration and teaching. If the last approach is followed, there is some duplication between the review of tests as they are used in the study of problems of curriculum, classification, methods, and the like, and the review of the study of these problems themselves.

The authors of the present number have chosen to place their emphasis chiefly on the third mode of approach, the review of the uses of tests. In addition they have reviewed the chief technical problems, particularly those which have an obvious practical bearing. They have made no attempt to review or evaluate the tests which are on the market, partly because of their overwhelming number, and partly because of the absence of adequate evidence on which to evaluate them.

This treatment of tests must be based partly on opinion, and involves, in some cases, differences of opinion. For example, the view expressed in the first chapter that the discriminative value of an item is to be measured by the correlation between the responses on that item and the responses on the test as a whole, assumes that the test measures a homogeneous, unified ability. Not all test makers agree with this assumption, as is shown in the second chapter. The reader will doubtless be interested to discover other differences in points of view. Such differences are inevitable where the discussion is based in part on opinion; and in this problem of the application of tests, on which there are now many questionings, it is inevitable that opinion shall be drawn upon. When the opinions are expressed by persons of such authority as those who have prepared this number they are bound to receive earnest consideration.

FRANK N. FREEMAN,
Chairman of the Editorial Board.

INTRODUCTION

THIS number of the *Review of Educational Research*, which has finally been completed only after many prayers and much anguish, departs somewhat from the style of the preceding numbers in that more space is given to the expressions of opinions of the authors of the various chapters and to the presentations of somewhat controversial issues. After considerable preliminary study of the task set for us during which was noted the vast literature on the subject including previous summaries and reviews, the committee finally decided that the most promising utilization of the space available would consist in directing discussions to *recent developments* and to *critical issues* as well as to the general review of the field. As chairman, I have deliberately stimulated the discussion of controversial issues in addition to more technical phases of testing and test construction. For it must be obvious that it is only through bringing out and constantly re-defining controversial issues that progress may be hoped for. It has seemed fitting also that essential technical implications of the measurement movement should be here treated in-so-far as space permits.

The reader will note considerable querulousness and some overlapping of topics, as the different writers belabor their various problems. Thus, for example, while Chapter V is concerned essentially with *uses of tests*, the other chapters also treat in some measure this same topic in the development of the various arguments, since the general theme is "Educational Tests and *Their Uses*."

As chairman of the committee I wish to express my appreciation of the excellent cooperation of all the members, the patience of Dr. Freeman, Chairman, Editorial Board, in allowing us extra time, and to acknowledge my indebtedness to my colleagues here in our Bureau of Research: Dr. Angela M. Broening, Dr. Grace A. Kramer, and Dr. Harold B. Chapman, without whose help this undertaking would have been impossible.

JOHN L. STENQUIST, *Chairman,*
Committee on Educational Tests and Their Uses.

CHAPTER I

Basic Considerations

THE most outstanding development in the field of educational tests during the last few years has not been in the technical advances in test construction, important as these advances have been, but rather in the major strategy as opposed to the minor tactics of the use of educational measurements.

Need of Measuring Growth over Period of Years

Limitations of the snapshot theory of testing—The customary conception of tests as being essentially snapshot affairs to be given, scored, and acted upon at particular moments and then forgotten, is now thoroughly discredited in all informed circles. Indeed, the snapshot theory of tests and the correlative methods of constructing and using them are difficult to account for as elements of the objective testing movement. It is only by an appeal to history that they can be made to fit into the picture at all. When viewed from a historical viewpoint, it is readily seen that they are survivals from the traditional "system" of subjective examinations that prevailed almost unchallenged until the emergence of the objective testing movement under the leadership of Thorndike, Terman, and their followers, and that had for its major purpose the "passing" or "flunking" of students, and the "enforcement" of exceedingly vague and variable "standards." The persistence of the snapshot theory and the correlative sporadic and unsystematic use of objective tests during the last decade, is evidenced by the fact that, with two or three outstanding exceptions, most of the so-called "standardized" tests available up to the year 1932 have existed in only two "equivalent" forms, which in some instances, at least, have turned out to be only "approximately" equivalent. We have had several series of so-called intelligence tests and several series of achievement tests in each of several matters for ten years or more; but even yet it may be confidently asserted that, with minor exceptions, no two such series have been made comparable, even though they have been edited by the same editor and published by the same publishing house. When viewed in the light of the insistence of the educational philosophers and psychologists on the importance of studies of growth, and of the insistence of the technicians on the importance of comparability in educational measurements, the situation just alluded to shows up the test-makers in an unsuspected light—a light tinged with a curious sort of conservatism and unconscious respect for tradition.

Systematic use of comparable tests—Fortunately, as indicated above, this conservatism has rapidly given way during the last few years to a clearer understanding of the basic necessity of measuring growth over a consider-

able number of years, and to a new and more adequate appreciation of the inescapable need for a systematic use of tests that yield comparable results. When considered in the light of these new conceptions, the prophetic and pioneering character of the twenty-odd comparable forms of the Thorndike Intelligence Examination for High-School Graduates, which were constructed more than a dozen years ago, and of the five comparable forms of the New Stanford Achievement Test, which became available in 1928, is clearly revealed.

One effort in the direction of providing comparable tests at the secondary and college levels has been made by the Cooperative Test Service (2, 3, 5, 7, 8). Practical programs in the systematic use of comparable tests and cumulative records (4, 9) have been developed in several states, including Minnesota, Wisconsin, Iowa, Ohio, Kentucky, Pennsylvania (1, 6) and others.

Improved Methods of Using Tests for Guidance

The impetus to this new conception of the rôle of tests has derived very largely from the hard school of trial and error. There are many schools that began the use of tests with enthusiasm and abandoned them in disillusionment. The cause of the collapse of tests in these schools is due to the disorganized piece-meal manner of using the test results, and to the lack of confidence inspired by "standardized" tests that are not only incapable of yielding comparable measures year after year, but are often very inferior in other respects. While there has been much room for technical improvement, the chief defect in the testing movement has been the neglect of building an adequate philosophy and system of using test results for effective and constructive educational guidance in the larger sense of that word. Not even the best tests have been well used. The test leaders have been slow to escape from the worst faults of the subjective system of examinations whose technical weaknesses they have so exhaustively analyzed and exposed. The greatest weakness in the traditional college entrance examinations, for example, is not in their technical defects, but in the snapshot, unsystematic, spasmodic way in which they have been used; in the systematic misuse which has ignored the real values and potentialities of the subjective examinations. It is not unexpected that teachers swamped by heterogeneous hordes of children, and that harrassed principals in high schools, and admission officers and deans in colleges, should be dominated by immediate expediencies to the point of failing to see the forest for the trees; but until recently too many leaders in testing and personnel work generally seemed to be dominated by the theory that test results and other personnel data were useful only in helping to solve the immediate exigencies of grading for credit purposes, of admission, of classification, of grouping for instructional purposes for a quarter-semester, or at most a year-period, with

variable and exceedingly opportunistic compromises with the curriculum. The idea that such data should be systematically recorded in comparable and meaningful terms, and that they should be made the subject of continuous and prayerful study and long-time planning on the part of the highest educational officers, as well as of the teachers throughout the whole educational ladder, has been slow to emerge, partly because the sporadic use of objective tests and snapshot interviews has given visibly better results and has thus served as a valuable stop-gap, and partly because of the persistent failure of all concerned to appreciate the magnitude of the guidance problem.

That the true nature and dimensions of the problem of guiding growing citizens through ten to twenty years of schooling into life in the wide world has been glimpsed only recently, is amply indicated by the fact that only a few years ago college personnel leaders apparently conceived of their tasks as being entirely independent of personnel and guidance work in the lower schools. The literature of a half-dozen years ago gives the impression that many college leaders were confident that high-pressure work with students *after* admission would solve the "college" personnel problem. Most teacher-training courses on personnel and guidance work are still announced as "Guidance in the Junior High School," "Senior High-School Guidance," and "College Personnel Technics."

As long as these arbitrary divisions of the educational ladder and the passage from one to another were exaggerated into an unnatural importance, and as long as the problems supposed to inhere particularly in the passage from high school to college were considered solvable by single sets of "Cross-the-Rubicon" examinations, it is understandable that the objective testing movement should have been first considered and judged primarily on the basis of its contributions to "college admissions" and other similar momentary crises up and down the educational ladder. That objective tests have acquitted themselves well when so used is amply indicated by the experience of a dozen years or more with intelligence, scholastic aptitude, and subjectmatter achievement tests. Indeed, it was the early success and patent superiority of objective over subjective tests when used for such purposes that prolonged the false hope that such momentary makeshifts as college entrance examinations and placement tests would suffice to solve the guidance problem, and thus delayed the concept of the rôle of comparable tests in the guidance problem, which concept now happily animates both testmakers and the increasing number of teachers and administrators throughout the educational ladder who have been active in the guidance movement.

According to this conception, ***the highest purpose and ultimate aim of the objective testing movement is not to make better college entrance or course-credit examinations, but to help inaugurate a continuous study of individuals throughout the whole educational ladder by means of systematically recorded comparable measures and observations which will make***

such spasmodic examinations largely unnecessary. From this viewpoint, college admission is merely one aspect of the larger and vastly more important total guidance problem. College admission will become an orderly and constructive process rather than the single nervous act which it now is—a part of the progressive learning and guiding of individuals into types of activities and ambitions which best suit their capacities, interests, and needs, from the kindergarten through the university.

Since the dawn of this view of the dimensions of the guidance problem and of the rôle of comparable tests in meeting the problem, many college leaders are beginning to ask questions which are a clear challenge to the leaders in the testing movement. "Why is it," they ask, "that after more than a decade of objective tests, the high schools are still unable to furnish the colleges with accurate and comparable indices of the achievements and abilities of candidates for admission to college?" Many enlightened secondary-school leaders are asking the same question regarding pupils that come to them from elementary schools in cities that have full fledged testing divisions, reference and research offices, and the like.

Few cities that have had test divisions for a decade can supply to the colleges cumulative records of comparable measures on their high-school graduates. The same is true of elementary schools regarding graduates to their own high schools. Plans for testing in such cities are highly variable, opportunistic, and, with rare exceptions, are not primarily designed to throw light on individual pupils as growing entities. The choice of tests is made without regard to the comparability of their results with those of tests used previously or to be used later, and no effective effort has yet been reported for making them comparable *ex post facto*. The reports are still made up largely of sterile comparisons of arbitrary or accidental groups of pupils called "classes" or "schools," and of "promotion" statistics which are nearly, if not quite, meaningless so far as the human engineering problem of educational guidance is concerned. The obvious implications of the presence in "eighth-grade" classes of pupils of "fourth year" achievement, and vice versa, are ignored almost as completely in the 1930 as in the 1920 reports of city "research" bureaus. The new conception of the rôle of comparable tests in the study of growing children has emerged none too soon to save the testing movement from obloquy and the schools from the wrath of the taxpayers.

Long-time Guidance of Individuals through Tests

The highest rule of measurement in education is not in the minor tactics of the classroom, but in the major strategy of educational guidance in the prophecy of long-term provisional goals for individual pupils, and the progressive modification of those goals in accordance with cumulative evidences of growth and of needs, intellectual, personal, and social. The usefulness of tests in diagnostic and remedial work is well established and

universally recognized; but in-so-far as this use of tests runs counter to, or does not contribute to, the long-time guidance of pupils, it is not to be accepted as an unmixed blessing. Some observers have feared, not without cause, that the diagnostic and remedial applications of tests are in some instances inspired (perhaps unconsciously) by the fallacious "leveling" implication of the democratic ideal of popular education, and thus may represent a subornation of test results to the enforcement of preconceived curriculums and "standards" which, in some instances at least, are wholly alien to the capacities and needs of the individual child who is the supposed beneficiary of remedial treatment.

The first question that the school should ask and answer at least provisionally several times each year is, "What *can* Johnny learn, and which of the things he can learn *should* the school, in the light of all the facts, try to help him learn?" Many of those who accept the fact of individual differences fear that this question, upon which systematically used comparable tests can throw so much light, is too often answered by reference to the curriculum book and to the administrator's convenience, and that the exigencies of maintaining the correctness of this pre-destined answer are too often the cue for diagnostic testing and "remedial teaching."

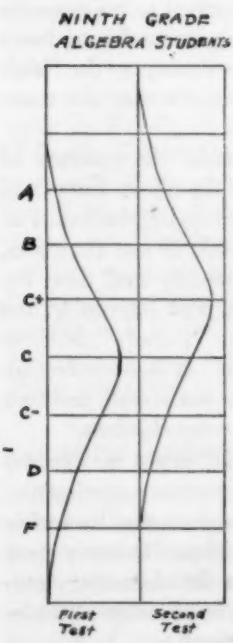
Tests should first of all tell *what* a pupil should *try* to learn—not *how* he may be cajoled, persuaded, or insidiously coerced into learning item x in the "standard" curriculum for grade n. If a pupil has difficulty in learning item x, this fact in itself may be evidence that x is not suited to his capacity and needs, and that he should, therefore, be given opportunity to learn something else, not forced by "remedial" treatment to live up to the "high and ever higher standards" so often found in perorations that are more impressive than meaningful.

Even if some do learn the prescribed minimum under the pressure of "remedial" treatment, the results might not be worth the effort. Indeed, if we consider the attitudes of despair, the feelings of inferiority, the habits of dependence, the frequently temporary and superficial, if not fictitious, character of forced learning, and the loss of opportunity and time for learning something that is within the comprehension and interest of the pupil, it is not by any means certain that the efforts to "remedy" children up to prescribed minimum are not positively harmful. It is no reflection on the curriculum makers to say that the curriculum is not sacred, and that the "prescribed minimum" may be a golden calf for *some* children.

Pupils who cannot learn after a year or two of trial might be excused even from some of the supposed "essentials" of the elementary curriculum, and encouraged to study some vocational subject or subjects that lie within their capacities and interests and thus have meaning to them. In many cases this procedure would not lessen their achievements in the elementary curriculum, and might engender better habits and more responsible attitudes in place of the disappointments that now frequently result.

The statistical "success" of certain "methods" involving "testing and remedial technics" has sometimes earned credit of a very dubious sort for the testing movement, and has frequently obscured and ignored or begged the fundamental guidance problem which has been emphasized in preceding paragraphs. The fact that certain testing and remedial technics have raised the average of ten thousand children to the extent of .5 sigma above the average of a control group may be entirely adequate evidence to prove the "superiority" of such technics over the methods used in the control groups for raising the average of such groups in the function or functions measured by the test; but it offers no evidence on the prior and more fundamental question as to which, if any, of the children involved should or should not attempt to master or get a smattering of the subject connoted by the tests used.

Let it be assumed, for example, that in a certain large area a certain change in tests, examination technics, or methods has indubitably raised the average of fifty thousand ninth graders in algebra (almost any other subject would serve in this example) to the extent of .5 sigma. Such gains have, in some instances at least, been interpreted to be analogous to raising the basal wage-rate of fifty thousand employees from, let us say, five dollars to five dollars and fifty cents per day. The inference is that algebra (plus the device or devices in question) is not only better, but better for *all* the pupils involved by some amount proportional to .5 sigma. The fallacy of this inference involves a great deal more than a statistical fallacy.



Indeed, for the purposes of this discussion, the statistical fallacy may be ignored; and attention be confined to the more important facts in the case, which are the fundamental assumptions and prejudices which underlie and motivate the fallacious inference.

What actually has happened in such cases here illustrated is not that all pupils have been raised to a degree of achievement which makes the learning worthwhile from a utilitarian viewpoint (assuming that any considerable number would ever have occasion to use algebra), but simply that the whole distribution of scores has been pushed upward, without changing its form to any appreciable extent. The nature of this "gain" is illustrated by the accompanying sketch, in which, for the sake of brevity, it has been assumed that each letter-grade group has been raised to the next higher group. If it is supposed that algebra in

any degree is good for all who now take the course, and that any degree of increased achievement in algebra is good for any and all such pupils, then the upward shift of the distribution might be regarded as a clear gain. But if all or part of this assumption is rejected, then the unmixed value of the gain is at least questionable. If the more reasonable assumption that algebra is not necessary for many is made and is (all things considered) undesirable or futile for some, and positively harmful to a few, the raising of the whole distribution may, to some extent, represent a misapplication of educational funds, a waste of children's time, and a reduction of their opportunities for genuine gains in other subjects and activities. The fundamental questions here are: (1) Which children are ever going to use algebra, directly or indirectly, sufficiently to give them an adequate return on their investment? (2) At what point on the achievement scale does a subject become worthy of its cost in a utilitarian or cultural or disciplinary sense? For the purposes of this discussion let it be assumed that the A's who have been raised to A+ have enjoyed a real gain, although it is not impossible that some of them might have gained something of more permanent value to their lives and to society by learning or doing something else in or outside the present curriculum. The same assumption might be made with regard to the B's or even the C's; but there are few unprejudiced persons who are familiar with the attitudes and achievements of the D and F groups in the average school who would seriously maintain that raising F's to D status is worth the time and effort involved. Indeed, there are many observers who would maintain that the raising of F's to D status represents a loss, on the theory that the traveller who goes farther on the wrong road is less fortunate than the traveller who has gone astray only a little way.

The above observations lead to the proposition that it is the highest function of the testing movement to help remove such fundamental questions from the realm of assumptions, which are complicated by academic log-rolling and pietistic prejudices, and to put the matter on a plane of rational consideration. The prime requisite for answering this question for individual children is to learn more than is now known about the capacities and interests of individual children. Ten years of research have fairly established the fact that while the occasional use of tests may mitigate maladjustments, such use of tests has signally failed to solve the problem of individual adjustment to the satisfaction of the schools or of society. Hence the plea herein is made for the systematic use of comparable tests and the careful study of their cumulative indications along with other relevant data for the long-time guidance of individuals and the comprehensive planning of their whole educational careers.

It seems clear that whatever contribution the systematic use of comparable tests may make to the problem of individual pupil adjustment will also tend to clarify, if not solve, some of the other problems that now beset our schools. In the hypothetical case considered above, it seems obvious that limiting algebra instruction to the smaller number of pupils who can

really learn and use or enjoy it, would have two or three effects of fundamental importance in addition to that of raising the level of achievement. In the first place, it would free a large number of children from efforts that are foredoomed to failure and allow them the possibility of devoting their time and energy to activities that are appropriate to their capacities, interests, and needs. The importance of this aspect of the matter has usually been overlooked by special pleaders for the curricular *status quo*. The usual form of their argument is that even those who fail are getting something out of it. Logically this argument is an appeal to the doctrine of despair; and unless its fallaciousness and harmful results are clearly apprehended, it will contribute to the continuance of the present wholesale sacrifice of children on the altar of erroneous prescriptions.

In the second place, the limitation of algebra instruction to the pupils described above would tend to raise the level of teaching ability, because fewer teachers would be required for the smaller number of children. The admittedly large number of persons who now teach algebra as an assignment, who do not understand its uses or appreciate it as a method of thought, and who have no vital interest in it, would be released for school tasks more appropriate to their capacities and tastes.

In the third place, this policy would reduce the cost of education and increase its constructive output. This does not mean that educational budgets would be reduced, but rather that they would be redistributed and more effectively used; and would thus open up the possibility of making the now hard-pressed and dissatisfied taxpayer willing to contribute larger appropriations.

Illustrative Use of Cumulative Records of Comparable Measurements

The importance of cumulative records of comparable measurements may be illustrated by the five-year records of two thirteen-year-old boys who were brought to the writer's attention in May, 1929, when they were both in the sixth grade. The principal had just administered the Stanford Achievement Test, Form A, and found that John had a total score at the 70th percentile, and Frank, at the 40th percentile of thirteen-year-old boys. The principal concluded that the difference was not great enough to justify drastic action; and in view of the unreliability of even the best tests and the almost total absence of any other meaningful information about the two boys, his decision to allow them to go "through the mill" seemed wise. Then it was accidentally discovered that both boys had taken Form B in May, 1928, and Form A in April, 1927, and two intelligence tests in September, 1926, and January, 1927. When the results of all these tests were graphed on the American Council on Education Cumulative Record Form, it was apparent that the two boys were in entirely different intellectual levels, their nearest approach to each other having been in their age-percentile ranks in May, 1929. In all the other tests John had consistently been at or above the 90th

percentile, and Frank, at or below the 30th percentile. With this array of evidence, the principal felt justified, in spite of the fact that they had received average teacher's marks of B and C, to make radical changes in their curriculums, and still more drastic changes in the provisional plans for their later educational careers. The tests given to these boys in 1930 and 1931 confirmed the indications of the earlier tests, with which they were comparable.

Importance of comparability—The importance of using tests that yield comparable measurements over a series of years is well illustrated by the two sample cases just described. Although the two boys were approximately two standard deviations apart, they nevertheless received average teachers' marks of B and C respectively. In a few cases the English grade of the less competent of the two was superior to the English grade of the more capable of the two boys.

It has already been pointed out that the fundamental weakness in the objective testing movement to date is that only two or three of the vast number of series of tests that have been available during the last decade have been made to yield comparable results. It is this major weakness in the history of test construction that shows up the existence of the snapshot theory of educational tests and the lack of an adequate appreciation of the systematic use of the tests in the study of growth in individuals. It is gratifying to be able to record that leaders of several of the better organized state testing programs are now definitely addressing themselves to the problem of constructing comparable examinations and using them systematically. As noted above, it was an appreciation of the needs for comparable tests that specifically led to the organization of the Cooperative Test Service of the American Council on Education.

The fallacies of standards—The fundamental weakness of standards has been their vagueness or meaninglessness. As currently used, the word *standard* has no place in educational literature outside the perorations of convention orators. The absurdity of the connotation in which it is used in such speeches is clearly exposed when the statements concerning educational standards are transferred to the more tangible rubrics like height and weight. The typical exhortation to teachers to seek "high and ever higher standards" is about as meaningful as would be the exhortations from doctors that children should be grown taller and ever taller. In this discussion it is not necessary to advert to the logical and psychological fallacies that are implied by such use of the word *standard*. It is enough to point out its meaninglessness.

Speaking more constructively, it is sufficient to point out that educational standards are necessarily individual, and in their fundamental nature are akin to the standards of tailors and shoemakers who judge the quality of their products by how well they fit the individual for whom they are intended and who pays for them, and how long they serve him.

Critical Issues in the Construction of a General Achievement Test[1]

The need for authoritative and specific descriptions of subjectmatter—In order to construct a valid and acceptable general achievement test, the test constructor must have before him an *authoritative* and *specific* description of the subjectmatter to be tested. Descriptions which have either one, but not both, of these characteristics are almost equally valueless. An achievement test is a collection of specific items dealing with specific materials. No single test item can test directly and by itself for the attainment of an ultimate objective. Ultimate objectives only represent convenient ways of describing immediate and more specific objectives collectively, and are tested only indirectly by testing for the attainment of these specific objectives. The materials used in testing for the attainment of immediate objectives are identical with, or highly similar to, the materials used in teaching to attain those objectives. Someone, either the test builder or the curriculum maker, must break down the ultimate objectives into immediate and specific objectives, and must prepare specific instructional materials for them, before construction of test items becomes at all possible. It is of little value for the test builder to know, for instance, that the ultimate objective of the social studies is "to develop better citizens, and well-rounded and cultured individuals," and that the intermediate objective is "to develop in the pupil a reasoned understanding of an insight into the nature, function, and evolution of . . . . institutions and practices and problems," unless, or until, he is told what *specific* items of related information, what *specific* relationships, and what *specific* ideas, generalizations, and broad concepts, constitute the accepted subjectmatter of instruction. Neither is it of much value to him to receive such suggestions for the content of test items, unless he has some reason to believe that this content will be generally accepted as belonging to the subjectmatter involved, that is, unless it is described *authoritatively*. The only authoritative sources of such specific descriptions of content which have been made available thus far to the test constructor are the textbooks and courses of study now being used in instruction. It is inevitable, therefore, that present achievement tests will follow closely the content of the best of present textbooks and courses of study, not because this content is that which ought (presumably) to be taught, but rather because it is the only content which thus far has been described in a form sufficiently specific, meaningful, and authoritative to make test construction possible.

Selection of the content of an achievement test—There are two conflicting considerations involved in the selection of the content basis for achievement test construction. In the first place, it is desirable that such tests avoid the appearance of encouraging the continuance of the *status quo* in curricu-

[1] For a fuller treatment of these issues, see: Lindquist, E. F., and Anderson, H. R. *The Nature and Function of a General Achievement Test in the Social Studies.* Iowa City, Ia.: the Authors (State University of Iowa).

lum content, thereby seemingly constituting a hindrance to progress and improvement in curriculum building. From this point of view, it would appear that tests intended for wide-spread use should be based on the most advanced and approved content available. In the second place, it is desirable that such tests provide as accurate measures as possible of the extent to which the pupil has achieved what he has been encouraged and given a reasonable opportunity to achieve. From this point of view, it seems that standardized achievement tests should be based on what is now being taught and learned. A compromise between these two conclusions is, of course, desirable, but it is important that both viewpoints be given adequate consideration in arriving at that compromise.

From the viewpoint of the teacher and the pupil it would appear desirable to construct standardized general achievement tests, so that they are as purely as possible measures of the success with which pupils have learned that which they have been encouraged and given a reasonable opportunity to learn, and so that they are as little as possible measures of the conformity of local courses of study to the latest suggestions for curriculum revision and improvement. While this condition could be approached by making the content basis of the test that of the typical local course of study, or by limiting it to that content which is common to the various local courses of study, such practice might appear to encourage or to sanction a static condition in the curriculum, and therefore is not to be considered. There is, however, a better alternative. That alternative is to base the test on that subjectmatter which is common to the *best* of the textbooks and courses of study now in use, and to place the major emphasis in the test on *reasoned understanding* of that content rather than upon the factual content itself.

Validity of test item—Any single achievement test obviously cannot hold the pupil directly responsible for an understanding of, or for the ability to use, or even for a verbal learning of, *all* of the specific information, relationships, ideas, generalizations, etc., which constitute a given field of subjectmatter. The subjectmatter of United States history, for example, consists of thousands of items of information, and thousands of related ideas, generalizations, inferences, and implications, based on that information, which it might be considered desirable for the pupil to learn and understand. If each of these elements could be given a weight proportional to its importance of value, then the pupil's total achievement or "general achievement" would be measured by the weighted sum of such elements as he has learned and understood. This concept of a true measure of general achievement, of course, can be only hypothetical. No single test, which will measure each of these many elements directly and individually in separate test items can be constructed and administered.

The items constituting any given general achievement test must be considered as representing only a very restricted sampling selected from all of

the items that might be constructed on the basis of the subjectmatter involved. Few schools will use a general achievement test that requires more than two hours for its administration, and tests of even shorter length are most in demand. Such tests cannot well include more than one hundred to two hundred items. With so restricted a sampling, it is highly important that each element in this sampling, or each item in the test, contribute as much as possible to the validity of the whole test. The validity of the whole test depends upon the degree to which it ranks pupils in the order of their true total achievement. It clearly follows that the validity of any single item in the test also must depend (within limits) upon the degree to which that item *of itself* discriminates between pupils of inferior and superior total achievement. If the ability of the pupils to respond correctly to a given item shows no relationship to their general achievement, that is, if the pupils who succeed on the item are not superior in general achievement to those who fail on the item—then that item cannot contribute to the central purpose of a general achievement test and cannot be defended for inclusion in the test, regardless of the "validity" of the item for inclusion in the course of study and regardless of its difficulty.

It often happens that two objective test items may prove equally "difficult" and may hold the pupils responsible for equally valid content from the curriculum viewpoint, and yet the actual responses made to one may be much more highly related to general achievement than those made to the other. Certain items, apart from their difficulty or desirability, represent far more crucial tests or indicators of general achievement than others. For example, in the field of elementary-school spelling, it has been shown that the ability to spell the word "adequately" is very highly related to general spelling ability, while the word "advisers" is misspelled more frequently by superior than by inferior spellers; yet both words are mispelled by the same proportion of pupils in a random sample of eighth graders, and both words appear equally desirable for inclusion in the course of study in spelling.

The worth or effectiveness of a test item depends, therefore, not only upon its desirability for inclusion in the curriculum and upon its "difficulty," but also upon its power to discriminate between pupils of high and low levels of general achievement. It is important to recognize this double aspect of the validity of a test item, and to see clearly the relation between "validity" from the curriculum viewpoint and "validity" for achievement testing purposes as determined by the discriminating power of an item.

The discriminating power of a single test item—The discriminating power of a single test item refers to the degree to which success or failure on that item by itself indicates possession of the general ability which is being measured. In relation to tests of the general achievement type, it may be defined as the accuracy with which a pupil can be placed along the general achievement scale on the basis of success or failure on the given item. An

item may be said to be perfect in discriminating power when every pupil who responds correctly to the item ranks higher on the general achievement scale than any pupil who fails on the item. An item may be said to have zero discriminating power when there is no systematic difference between the general achievement of the pupils who succeed on the item and those who fail.

Otherwise stated, an item is said to discriminate if the pupils who respond correctly to that item are, on the average, superior in general achievement to those who respond incorrectly. If the pupils who succeed on a given item are, on the average, just equal in general achievement to those who fail, then the item has no discriminating power. The degree of discriminating power of an item therefore depends upon the magnitude of the difference between the *averages* in general achievement of those who succeed and those who fail on the item.

Various hypothetical degrees of discriminating power for a test item of 50 percent difficulty are represented in Figure 1. This figure shows the various types of relationships which may be found between general achievement, as measured by a comprehensive criterion test, and the ability to respond correctly to a single given item (in this case an item answered correctly by 50 percent of the pupils in an experimental group). The vertical scale in this figure indicates the percent of pupils who responded correctly to the item. The placement of pupils along the (horizontal) general achievement scale is determined on the basis of their percentile standing on the criterion test. The "line of discrimination" for a given item indicates the percentage of pupils at each level of general achievement who responded correctly to the item.

Line MM in Figure 1 represents the line of discrimination for an item (of 50 percent difficulty) which shows perfect discriminating power, since every pupil below the 50th percentile of general achievement missed the item, and every pupil above the 50th percentile succeeded on it. The pupil who responds correctly to this item may then be accurately placed on the general achievement scale with reference to one point, in this case the point of median achievement. Only a dichotomous classification, of course, is possible—the pupil's distance above or below the median point in general achievement cannot be determined on the basis of this item alone.

Line UU in Figure 1 represents the line of discrimination for an item (of 50 percent difficulty) which has zero discriminating power, since the same percent of pupils at every achievement level responded correctly to the item. When a pupil responds correctly to an item of this type, there is no greater reason for placing him on the lower part of the general achievement scale than the upper; that is, he is no more likely to be good than poor in general achievement. Items of this type have no functional value in a general achievement test, regardless of their other characteristics.

Line VV in Figure 1 represents the line of discrimination for an item (of 50 percent difficulty) which has negative discriminating power, since it is answered correctly more frequently by pupils of inferior general achievement than by pupils of superior achievement. A pupil who responds correctly to an item of this type is more likely to be *low* in general achievement than one who responds incorrectly. Such items can have no functional value in a general achievement test, unless the pupils are "given credit" for making the *wrong* response, which obviously is impracticable. A large number of items of this type has been discovered by the authors in their experimental try-outs of test materials in the social studies, and concrete illustrations of them will be presented later.

Between the extremes of perfect positive and negative discriminating power all degrees of discrimination may be found. These are illustrated (for items of 50 percent difficulty only) in Figure 1 by lines NN, OO, PP, QQ, etc.

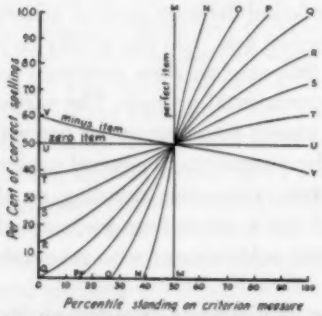


Fig. 1 Hypothetical Lines of Discrimination Showing Stages Between Perfect and Minus Discriminating Power for Items of Fifty Per Cent Difficulty

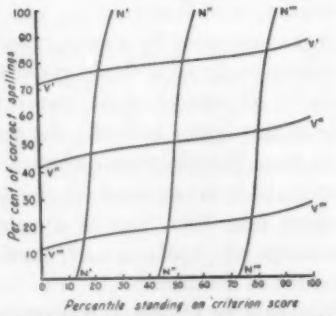


Fig. 2 Hypothetical Lines of Discrimination Illustrating How Items of Various Difficulties May Be Either High or Low in Discriminating Power

It is very important to note that there is no apparent reason for assuming any relationship between the discriminating power of a test item and its "difficulty," or percentage, of incorrect responses. For example, while only ten out of one hundred pupils may succeed on a given test item, it may happen that these ten pupils are on the average no higher in *general* achievement than the ninety who fail on the item in question. Similarly, an item may be answered correctly by eighty out of one hundred pupils (that is, it may be an "easy" item), and yet among the twenty pupils who failed on this specific item there may be many who are superior in general achievement to some of those who succeeded on it. The difference in average achievement between those pupils who fail and those who succeed may be much greater for one item than for another of the same difficulty, or for a given "easy" item than for another that is very difficult. An item of any difficulty may have any degree of discriminating power.

In Figure 2, hypothetical lines of discrimination are illustrated for good and poor items at each of three levels of difficulty.

Line N′N′ in Figure 2 represents an item of 80 percent correct responses (20 percent difficulty) with very high discriminating power, and line V′V′ represents an item of the same difficulty but with very low discriminating power. The lines N″N″ and V″V″ represent items of 50 percent difficulty with high and low discriminating power respectively. Lines N′N′, N″N″, and N‴N‴ represent items of marked differences in difficulty but with the same degrees of discriminating power. Pupils may not always be assumed to be high in general achievement simply because they succeed on a "difficult" item, or low in achievement simply because they fail on an "easy" item. There are many "difficult" items which are more frequently answered correctly by inferior than by superior pupils, and many "easy" items which are more frequently missed by good than by poor pupils.

In terms of this illustration, it should be apparent that the ideal test would consist of items of high discriminating power distributed evenly over the difficulty scale. In other words, the ideal test would contain some easy items that discriminated sharply at low levels of achievement, others of medium difficulty that discriminated sharply at high levels of achievement. The significance of this ideal distribution of item difficulty will be discussed later. Statistical technics have been developed for expressing the discriminating power of a test item in terms of a single numerical index, or "index of discrimination."

Summary of restrictions upon the content of a general achievement test—In summary, then, the specific content of a general achievement test intended for a given group of pupils is subject to the following definite restrictions:

(1) If we think of the entire group of pupils as separated into a number of levels of general achievement, then for each of these levels the test must contain an approximately equal number of items calling for information that the pupils at that level *have* learned; or for ideas and generalizations that they *do* understand; or for judgments, applications, or reasoning of which pupils at that level *are* capable.

(2) The items thus selected with reference to each level of achievement must *discriminate* as sharply as possible between pupils above and below that level; that is, it must (ideally) be highly probable that all pupils above that level will succeed on each of those items and that all pupils below that level will fail on each of them. (This is particularly important in recognition tests with reference to items that call for judgments beyond the ability of the pupils tested.)

(3) The items must be such that the response scored as the "correct" or "best" response would be considered so by competent authorities. (If the wrong response of certain items could be scored as "correct," these items would meet the first two requirements.)

(4) The items must hold the pupil responsible for understandings, abilities, or information that it is believed will contribute to the realization of the objectives of instruction; that is, they must be based upon subjectmatter which has been *authoritatively* and *specifically* selected and described for purposes of instruction and which does belong to the field of subjectmatter involved.

These discussions of the technical aspects of test construction have shown that the problem of defining a field of subjectmatter in which general achievement is to be measured, and the problem of selecting elements of that subjectmatter for the construction of items in a general achievement test, present very marked and significant differences. The items in a test do not by any means represent a *random* or representative sample of the content of the field in question. While every attempt should be made to make the sampling as representative as possible, certain items must be excluded because of technical considerations. What the test may contain is very largely a function of what has been learned and of what is the level and range of achievement in the group to be tested. The content of achievement tests, therefore, cannot be expected to parallel the course of study exactly, and such examinations can neither be used to "check" nor be checked by course of study outlines directly. The problem of selecting test items cannot be left to the subjectmatter expert or to the subjective judgment of anyone but is mainly a technical problem, and must be based upon objective facts secured from actual trials of large numbers of items with pupils of the kind to which the completed test is to be administered.

CHAPTER II

The Selection of Test Items[1]

THE problem of the selection of test items has existed since the first test was given. In the days before the advent of standard tests the basis of selection was entirely analytical. The testers simply included those items or questions that, within the limits of the subjectmatter, seemed to them as of most importance. This was a matter of individual preference and of subjective judgment. The trend of the times was to allow much freedom to the individual examiner and everyone seems to have been happy.

Conditions of the sort that has just been described existed until the advent of the standardized test. The appearance of the standardized test marked the beginning of a new point of view concerning the selection of test items. Since then this selection has been based more or less on statistical considerations. Among the various workers the relative emphasis given to the analytical and statistical methods of approach has varied. Some workers have favored the analytical approach almost exclusively, some have based their work mainly upon statistical considerations, while most have made some more or less balanced combination of the two. At all events a line of cleavage between analysis and statistics has existed and persisted since the advent of the first standardized test.

Analytical vs. Statistical Determination of Test Items

The appearance of the Stone Arithmetic Tests in 1908 (56) marked the beginning of the standardized test in its more modern sense. Stone used both analysis and statistics in the selection of his test items. He analyzed or divided his items into classes. Some of them related to computation and others to problems. The computation items represented each of the four fundamental processes. The problems were selected so as to afford situations equally concrete to all children in the A sixth grade. Large numbers, particular necessary requirements, catch problems, and all subjectmatter except whole numbers, common fractions, and United States money were excluded.

The statistical nature of Stone's work lay in the fact that he selected items of known difficulty and arranged them in scaled order. From this simple and apparently trivial beginning has come a great statistical movement in education—one which has threatened, at times, to crowd out entirely the analytical approach to the selection of test items.

In 1909, the year after the publication of the Stone Tests, Thorndike published his handwriting scale. The items in this scale were selected

[1] Bibliography prepared with the assistance of Dorothy Adkins.

entirely by statistical procedure. Apparently the entire object was to secure samples of handwriting of such a nature that their values in terms of quality would be approximately one P. E. apart after they were arranged in scale order. Analysis seems to have played almost no part at all.

The Thorndike Handwriting Scale is the pattern for all of the scales which have appeared since. The method used in the derivation of the scale has been applied in other subjects and in the rating of human beings. In all these subsequent applications statistical considerations have been paramount, but the analytical approach has also been in evidence. In English composition the first scale was published by Hillegas. Like the handwriting scale it was statistical in nature. Some of the paragraphs were frankly artificial while others were selected from the works of standard authors. Apparently any source was utilized to obtain a composition of the desired quality. The disadvantages of this procedure were evident and soon a supplement of the Hillegas Scale appeared (67). Trabue made a substantial improvement upon the Hillegas Scale when he restricted most of his compositions to the subject *What I Should Like to Do Next Saturday* but catered to statistical considerations by adding some compositions of a miscellaneous nature at the top of his scale.

Trabue's supplement was followed later by Thorndike's extension of the Hillegas Scale (63), in which there is definite evidence of analysis in the selection of the compositions. The Thorndike extension contains twenty-nine compositions, but only ten deal with the mere supplying of details. Three are concerned with cause and effect relations, three with advantages and disadvantages, five with interpretation, four with functional relations, one with inference, and one with argument. The remaining two compositions were letters. Since eight types of paragraphs were included, there was no homogeneity. To make matters even worse pupils were asked to write stories which had to be limited to narration only, and these stories were graded on the basis of a scale which contained mostly non-narrative material. Obviously statistical considerations were influential.

Ballou (12) saw the defects of this procedure. He criticised the Hillegas Scale on the ground that it "aims to measure too varied a product," and contains compositions that are "artificial, bookish, and not typical of good school work." He objected also to the fact that no conversation was included. Ballou said, "A scale should not measure too complex a product. To attempt to measure the several forms or types of English composition by one and the same scale is like trying to measure heat, light, and color by the same instrument. . . . The compositions of a scale must be analyzed as to merits and defects, also there is not only no guarantee but little possibility that the users of the scale will interpret the qualities of compositions any better than they now interpret the qualities of compositions without the use of any objective scale." The Harvard-Newton Scales were con-

structed to correct those faults. They contain separate scales for description, exposition, argumentation, and narration. With each composition there is a list of merits and defects and a comparison of that composition with others in the scale. The Harvard-Newton Scales are much better than those which preceded them. They have merited much more consideration than they have received.

Unfortunately Ballou failed to break through the statistical mode of the time. Seven years later when composition scales began to appear again, statistical considerations were still dominant. The Hudelson and Willing Scales show little evidence of analysis, but Lewis (33) analyzed his items into letters containing orders, letters of application, social letters, and simple narration. During the eleven years since Lewis published his scales, only one contribution of analytical nature seems to have appeared in English composition. That is the work of Odell (42). Odell's Scales were designed to rate pupils' answers to thought questions. There are nine scales covering analysis, cause-effect, comparison, criticism, discussion, explanation, relationship, argumentation, and summary. The compositions for each scale are motivated by an appropriate stimulus question, and each scale includes lists of merits and defects.

While Odell apparently had no intention of doing so, it is true nevertheless, that he has constructed a composition scale which, in the estimation of the reviewer at least, is the best that has yet appeared.

The work of Buckingham (16:13) raised another troublesome question. In the construction of his Spelling Scale, Buckingham selected words that "were sufficiently common in the speaking vocabulary of third-grade children," and whose spelling difficulty in most cases was "great enough to test the ability of eighth-grade children." In this selection he had regard for both the analytical and statistical points of view. The new problem centers, however, in his definition of difficulty. He assumed that the proper difficulty of words for each grade is that in which the word is spelled correctly by half of the pupils and missed by the other half. This conclusion was arrived at on the basis of considerations that are purely statistical in nature. Buckingham argued that words which all of the children of a given grade can spell and words which none of them can spell are alike useless for purposes of measurement. From these two theses he arrived at the conclusion that the midpoint, the one in which half of the student population answers correctly and half answers incorrectly, is the best *crucial* score for each grade.

In his criticism of the Buckingham Scale, Otis (44:795) said: "The most crucial words for any grade are believed to be those which may be spelled by approximately 50 percent of the pupils of that grade, or those words of approximately equal difficulty of which the average score of the pupils of that grade will be 50 percent." Otis also made a distinction be-

tween testing the ability of a pupil to spell words with which the child has had ample time to become familiar and the probable number of words in the language which the pupil can spell. If the former is desired, Otis endorsed the plan of Ayres (11) who placed the crucial score at 84 or at words which 84 percent of the children spell and 16 percent miss. On this point Ayres said (11:37):

> All of the scales have been arbitrarily cut off at 50 percent, partly because it is doubtful whether any useful teaching process is served by testing children on words of which they cannot spell more than 50 percent correctly, and partly because children of the lower grades attempting to spell difficult words frequently fail, not because of the inherent difficulty of the spelling, but because the word form is not yet definitely a part of the children's regular vocabulary.

Courtis (20, 21, 22) agreed with Ayres to the extent that he called the crucial grade from 76 to 80. Monroe placed this same grade at 70 percent and Haggerty, at 66 percent.

The issue here raised is fundamental. Shall pupils be measured against a sort of absolute social standard or in terms of what they have been taught? The statistician inclines to the former point of view and the teacher and school administrator, to the latter. The school administrator can advance arguments against Buckingham's position. As a pre-test it might be advantageous to know that none of the children are able to score even a single point on the performance of a task that is to be undertaken. In like manner there may be vast satisfaction to the teacher and all concerned when all of the pupils get perfect scores on some test. Surely no one would feel angry or disheartened when a group of pupils is able to score 100 percent in their responses to the fundamental combinations in arithmetic.

The source of the entire difficulty, as Monroe and Ayres indicated, lies in the failure *to consider tests in terms of the purpose with which they are used.* The statisticians have been interested in testing for purposes of classification. Under such conditions the 50-50 plan is desirable and defensible. In the measurement of achievement, however, the case is quite different. No one is satisfied with the 50-50 plan. Passing grades are seldom, if ever, set that low.

The distinction between testing for classification and for achievement has not been clearly seen by many workers in the field. Ashbaugh (10) used the 50-50 plan and the same procedure was followed by Cook (18), Gates (27), Kelley (29), Ruch (50), Symonds (57), Thorndike (60) and T. Thurstone (64). All of these workers were interested mainly in improving the statistical method as applied to testing. To them, testing is a measuring and not a teaching device. As McCall (37) put it they are "willing to sacrifice diagnostic ability to statistical beauty." Wood (77) said: "In the construction and administration of examinations measurement must not be confused with pedagogy." Apparently the test maker is to be justified in going outside of the school curriculum for a 50-50 test item. The fact

that the desired item is one that the child has not had an opportunity to learn in school is in no wise a deterrent to the statistically minded test maker. His search for 50-50 items must not be limited by the content or objectives of a course of study.

Against this point of view Monroe (39) and Wilson (75) protested. Wilson's protest was answered by Kelley (29). Portions of Kelley's defense are worth quoting. He said: "The merit of a yardstick used in measuring the height of children is not judged by its effect upon the heights of the children measured. That a yardstick does not affect the height of a child is not usually held against it."

Kelley admitted that "to best test the specific advances of pupils that have been subjected to a particular instruction in spelling, it is necessary that the test involve the exact words used in the spelling lessons"; but added, "The authors of the Stanford Achievement Tests believed that their function was different from this; and that they were called upon not only to test pupils upon specific words studied, but to test for a spelling ability circumscribed only by the field of common usage."

Kelley pointed out that school grades are not discrete. Some pupils in every grade have abilities that are from one to two grades higher while others really belong on levels one or two grades lower than the grade in which they are placed. Consequently if children in the eighth grade are to be tested adequately the words (items) in the test must vary in difficulty from the sixth- to the tenth-grade levels. Since 50-50 words are desired, the test must include 50-50 words for the sixth, seventh, ninth, and tenth grades as well as for the eighth grade. The 50-50 words for the ninth and tenth grades must therefore, of necessity, be those that have not been taught. The Stanford Achievement Tests are, therefore, primarily an instrument for the classification of pupils. The measurement of achievement is a by-product.

The authors (31) of these tests distinguished between the measurement of the achievement of educational objectives and classification, but they combined both functions in the same tests and named their tests in terms of their minor function. This tends rather to perpetuate the confusion which exists among the rank and file upon this point.

Difficulty of Test Items

There is yet another open question of long standing with reference to the selection of test items. Previous to the appearance of the Stone Tests (56) it was generally assumed that test items are of equal difficulty throughout a given test. The essay questions of previous tests were usually presented in groups of ten or less, and the same credit was allowed for each question correctly answered. Stone, however, arranged his items in order of difficulty. Some of them were more difficult than others. The scale makers followed this procedure also. Courtis (21, 22), however, adhered to the

older policy. The items in the Courtis Tests are all of approximately equal difficulty. Woody (78) followed in the steps of Stone and got into difficulty because he could not always find an item which was one unit more difficult than the preceding one. In such a situation he preferred to bring in a non-essential item of the proper difficulty even when it was necessary to omit samplings of essential points in arithmetic. Monroe (39) criticised this procedure and asserted that test items should be restricted to the list obtained from the analysis of subjectmatter. Monroe also attacked the policy of scaling items as a whole. Such scaling presupposes a gradual increase of ability on the part of the pupil; whereas the teaching of a given item should result in an abrupt increase. This is in line with the general assumption that pupils know nothing at all about given items before they are taught, but have virtually perfect knowledge of them after teaching.

This question has never been settled and as a result we still have both kinds of tests. The advent of the so-called "new form" examinations, such as the true-false, multiple-choice, and completion types, has perpetuated the policy of constructing tests out of items of approximately equal difficulty. At the same time there is a continual crop of standard tests and scales in which the items have increasing difficulty.

Validation of Test Items

The question of the difficulty of test items merits further consideration, but it is necessary to turn aside for a time to consider one more new element that the statistical method has brought with it. In 1914 Thorndike (61) published his first report on the measurement of ability in reading. In the vocabulary portion of the test Thorndike selected as test items names which belong in certain well-known classifications, and controlled the depth of comprehension by limiting definitions to the ability of the pupil to associate each species with its appropriate genus. This procedure constituted a "criterion" for the mastery of word meanings. The war brought the word *criterion* into more general use. It also tended to put test construction on a rather empirical basis. The test maker looked first for a list of items that would yield a "normal distribution" and would correlate as highly as possible with some criterion. A test constructed in this manner was considered "valid" without regard to the nature of the test items. The exigencies of war made it impossible to measure the activities of men on the actual battle front. Therefore an "indirect measure" was sought. Whenever it was impossible to measure a function directly, recourse was had to the measurement of another function that is highly correlated with the first. This "indirect measurement" was necessary then and may still be quite useful, but it has its disadvantages. It suffers in the first place because it has proved impossible to find functions that correlate

anywhere near perfectly with any criterion. Secondly, such procedure destroys all consideration of individual test items. This sort of measurement reduces analysis almost to the vanishing point. It represents statistical method at its maximum.

After the war the influence of this procedure upon achievement testing was marked. Validation became almost entirely a matter of correlation with a criterion. The test item was completely lost for six years until Vincent (73) discovered it again. In the meantime far from adequate attention was given to the selection of criteria. Upon this point Ruch and Stoddard's (51) excellent summary of methods of test validation threw some light. Of the twelve methods which Ruch and Stoddard mentioned five involve correlation with criteria. These criteria include judgments of competent persons, rating scales, school marks, previously validated measures, and tests of other intellectual, non-intellectual, or educational tests. All of these rest ultimately on subjective judgments. This is obviously unsatisfactory, but the reviewer has been able to find only a few studies in which anyone has tried to do anything about it. Toops and Royer (66) detailed a method for constructing a criterion. The treatment is highly technical, but well worth the study of all those who plan to use this method of test validation and of those who wish to improve methods of constructing criteria. Thorndike (62) based his criterion of "intellectualness" on certain objectively verified assumptions concerning that function. He (62:156) reported that "on the whole it is certain that we cannot trust any consensus of present opinion to provide an accurate measure [criterion] of the difficulty or of the intellectual difficulty of a single brief task." Gates (26) and Cook (18) were skeptical about criteria, but neither offered constructive suggestions concerning their improvement.

The rediscovery of the test item dates from the publication of Vincent's study (73) in 1924. Since that time there has been a growing feeling that test items should be weighted, not only in terms of their difficulty, but also with reference to their individual contribution to the correlation of the test with its criterion. The contribution of the individual item has been estimated in a variety of ways. Lentz, Hirshstein, and Finch (32) gave the most complete summary and evaluation of these methods. Cook (18) also gave an excellent treatment of them. A few of the methods are worthy of special mention here. The bi-serial R method stands high in the opinion of all but it involves much labor. For this reason most of the workers prefer some modification of the overlapping method. Cook (18) found that one form of the overlapping method is really superior to the bi-serial R. Lentz, Hirshstein, and Finch (32) found the "upper and lower third" method best. Clark (17) reported a method of obtaining an "index of validity" for a single test item by means of the formula I. V. $= \frac{P-D}{1-D}$. In this formula

I. V. is the index of validity, P the percentage of the criterion group failing to answer the item correctly, and D the percentage of the group who took the sub-test and failed to answer the same item correctly. An inspection of this formula shows that the right hand member reduces to zero when P = D and to infinity when D = 100 percent. But when P = D the index ought to be a maximum and it ought to be a minimum when D = 100 percent. It looks as though Clark's formula should have been written

$$\text{I. V.} = 1 - \frac{P-D}{1-D}.$$

Lindquist and Anderson (34) obtained an "index of goodness" for single test items in terms of the ratio of errors made by the upper and lower quartiles. They published a test in world history which shows an "index of goodness" for each test item. Similar information is available also in the study of W. Wilson, Welsh, and Gulliksen (76). Toops (65, 66) preferred to validate test items by means of the regression equation technic. His plan has the great advantage of preventing the repetition of test items which overlap in their contributions to the criterion. The disadvantage lies in the amount of statistical labor involved. More work of this sort would constitute a valuable contribution to our knowledge of the validity of single test items.

All of the statistical methods of validating test items that have been mentioned thus far have been applicable only to test items requiring dichotomous answers. The answer must be either right or wrong. This is obviously a disadvantage, but fortunately McCall (36) described a method that does not suffer this limitation. The method is statistical and technical. It suffers somewhat from the inadequate explanation which McCall's students gave concerning it.

The foregoing discussion of test items indicates that considerable progress has been made in the selection of test items, but even yet there is doubt as to whether the procedure is really worthwhile. Corey (19) and Whelden and Davies (74) reported in favor of weighting test items, while Douglass and Spencer (24), Odell (41), and Potthoff and Barnett (48) found that weighting is not worthwhile. Tyler (71) criticised the practice of validating individual items by their relationship to the total test score. He claimed that the items are not highly valid and that their relationship to the total test score assures mere homogeneity rather than validity. The doubt which arises concerning the use of a criterion in connection with the validation of test items has led some to pin their faith to reliability as the best means of validation. In this connection Symonds' article (58) and Kelley's note (30) are of interest.

Having traced the development of statistical method, it is now time to give further attention to the analytical method of selecting test items. It

has been pointed out that the original method was analytical and that at least some trace of the analytical method is present in nearly every statistically derived test. But the analytical method has never been entirely without its defenders. Previous to 1917 Ballou made a careful analysis of difficulties in common fractions, and found, for example, fourteen types of abilities involved in the addition of fractions, each of which called for a different specific ability. Monroe (39) in 1917 approved of this procedure and commended the Cleveland Survey Tests in Arithmetic because they were based upon a similar analysis of subjectmatter. In 1927 Monroe (38) wrote: "It may be interesting to ask students to respond to puzzles or other exercises which are not in agreement with accepted educational objectives, but it is seldom worthwhile and is to be condemned." Monroe also favors the retention of the essay examination as an essential means of testing the achievement of important educational objectives. Previously (1923) Monroe and Souders (40) listed fifty types of essay examination questions which were "related to the daily objectives of the teacher." Odell made further contribution to the analytical method of approach.[1] In 1922, Pressey (49) wrote: "A good test covers only the really important points of a subject. . . . The best tests are based on very careful research as to the fundamental objectives in the subject concerned, and the material is selected with reference to its importance for these objectives." Wilson (75) said: "The first fundamental criterion of a test should be that it serves the main curricular aim of the subject tested. The second fundamental criterion is that the test should properly reenforce good methods of teaching." Tyler (68, 69, 70, 71, 72) is a strong believer in the analytical method. He (68) criticised those who assumed that measures of information are adequate measures of ability to think. He reported that the correlation between information and certain types of thinking in college biology ranged from .40 to .46. The correlation between information and skill with the microscope was only .02. Tyler's fundamental thesis is that objective tests should be concerned with obtaining objective evidence of the degree to which students are attaining the important goals of education. "In the past there has been the common failure to distinguish between the content of a subject and the mental processes which a student of this subject is expected to exhibit." Tyler insists that those who build examinations in college subjects must have "training in the analysis of the psychological processes characteristic of college subjects." He opposes the use of a single test in a subject because each subject has more than one objective. A test is needed for each objective for which attainment is to be measured. "To ignore this is to put a straitjacket upon education."

---

[1] Odell, Charles W. *Traditional Examinations and New Type Tests.* New York: Century Co., 1928. 469 p.

This is a reflection of a growing view that tests may easily block educational progress. It is claimed that teachers are inclined to teach test items—that they do not discriminate between the items that are related to educational objectives and those that are present in the test for statistical reasons only. These fears are also responsible for the feeling on the part of some that all testing should be abolished. Possibly it was a fear of this sort that motivated Peters (46, 47) to study the relation of standardized tests to educational objectives with special reference to social needs. Peters found that twenty-two different types of test validation had been used by the test makers of 183 tests. Less than one-third of these had used analytical methods. Peters (47:150) concluded that "only a few tests rest upon a systematic survey of social needs." Other writers who indorse the analytical approach include Barr (13), Seashore (53) and Sones and Harry (55).

The Essay Type of Question

For sixteen years it has been customary for test makers to decry the essay examination because of its unreliability. The result of this opposition has been almost to destroy the instrument whose chief function is to measure a pupil's familiarity with the thought of our great thinkers. The unreliability of the essay examination is said to be the result of two things: subjectivity and limited sampling. Most of the workers in the test field have been predominantly statistically minded and to them a lack of reliability is naturally a fatal defect. As a result, the essay examination has simply been ignored. No attempt has been made to question the condemnation that it has received, and until recently no one has tried to improve this type of test. Osburn (43), however, reported a study in defense of the essay examination. He found that the subjectivity of scoring can be decreased markedly by providing a list of acceptable answers, care in the statement of questions, and control of the amount to be taken off for formal errors. With these improvements Osburn obtained scores of high reliability. For example, the mean of the scores on "Name the five most famous Spanish explorers and their explorations," assigned by seventy-five relatively untrained scorers was 34 times its semi-interquartile range. The corresponding critical ratio for "What are the functions of leaves?" was 5, and that for "Compare the Articles of Confederation with the Constitution" was more than 6. For twenty-nine questions in history and biology all but five scorings showed a critical ratio of more than 3. While these results are inconclusive, they show that the subjectivity of the scores of essay examinations can be controlled to a very large extent.

The criticism of the essay examination from the point of view of inadequate sampling is usually the result of superficial consideration only. A number of writers have compared the five or ten questions of the essay examinations with the one hundred items of a true-false or other new-form

examinations and have concluded at once that the essay examination includes very little sampling. Such writers fail to consider the elements in the answers to essay questions. These are the only items of an essay examination that can be compared legitimately with the items in a "new-form" test. The average number of items in the answers to essay examinations runs around nine or ten per question.

A few writers such as Eurich (25), Sims (54), Russell (52), Talbott and Ruch (59) have examined essay questions more closely. Both Eurich and Sims were interested in making essay examination questions covering the same ground as "new-form" tests. Sims used unimproved essay questions, and Eurich's results are somewhat in doubt because the tests which he compared contain different numbers of items. He has sixty-nine items in his essay test, ninety-four in completion form, thirty-nine in multiple choice, and fifty in true-false. It is difficult to understand how a legitimate comparison of types of examinations can be made when the number of items in each test varies so greatly.

Russell (52) and Talbott and Ruch (59) were interested in comparing essay and "new-form" questions on the basis of intensive versus extensive sampling. Russell presented a diagram to show that a given pupil may get zero or 100 percent according to how the essay questions are selected. Talbott and Ruch concluded that "on the average the essay question calls forth less than half of the pupil's knowledge." All of this sounds conclusive, but Osburn (43) questioned the soundness of the extensive sampling procedure. He pointed out that the theory of sampling, as applied in education, is borrowed from the field of mathematics. In mathematics this theory involves two assumptions: homogeneity and chance distribution of the content to be sampled. In education, Osburn pointed out, the content is neither homogeneous nor subject to the law of chance distribution except within such limited fields as may be covered by a single essay question. If these statements are true, serious doubt is cast upon all tests such as those of the "new-form" type that are based on extensive sampling, and the foundation of most statistical methods of validating test items will be dangerously beset.

Finally the new interest in the essay examination has brought out some evidence that the old-time essay examination was not as bad as it seemed. Brinkley (15) found that "new type tests prepared by a group of high-school teachers gave, on the whole, lower validity coefficients than old type tests." Even after instruction had been given in formulating new type tests the results as to validity were "slightly poorer." Brinkley succeeded in making new type tests that showed slightly greater validity than the old type, but "the significance of the difference could not be determined." On the whole Brinkley seems quite favorable to the analytical approach. Gates (28) used essay examination scores as an important element in his

criterion for true-false tests. Bayles and Bedell (14) found that completion in story form (mutilated essay examination answers) is more valid than modified true-false, matching, multiple choice, and ordinary true-false forms when the same unit of subjectmatter is concerned.

Summary and Conclusions

The foregoing survey of the literature relating to the selection of test items makes no pretense to completeness, but it will be a matter of regret if any of the major contributions have been overlooked. The aim has been to raise into the clear as many as possible of the crucial problems that are involved and present them as clearly and fairly as possible. It will be noticed that the conflict between the analytical and statistical point of view has been prominent throughout the history of the test movement. This conflict is increasing rather than decreasing at the present time. It is impossible to see what the outcome will be, but in all probability some sort of compromise will result eventually. In the interests of more complete harmony, it might be well to distinguish between testing and measuring, as has been the case in chemistry. In that subject, testing is a qualitative procedure while measuring is quantitative. The former is concerned with the mere detection of the presence or absence of a given ingredient, while the latter is concerned with *how much* of the ingredient is present. Possibly the analysts will eventually be content to restrict themselves to testing as thus defined, leaving measurement to the statisticians.

Other crucial questions relate to whether or not all test items should be of equal value, whether or not these items should be related to the curriculum as it now is, whether or not the extensive theory of sampling is sound in education, and how to improve the criteria of validation.

CHAPTER III

Recent Developments in Statistical Procedures

IMPORTANT developments in the applications of statistical methods to test construction date, with few exceptions, from about 1920. A few sources prior to 1920 should receive specific mention because of their great influence on later developments. The publication in 1913 of Thorndike's text in statistical method (231) first drew the attention of American educators to quantitative thinking. Kelley's doctoral dissertation (147) in 1914 introduced the technic of partial correlation and multiple regression to test workers. The years 1916 and 1917 saw the publication of three influential books: those of Starch (221), Rugg (209), and Monroe, DeVoss, and Kelly (171). To these sources may well be added Whipple's *Manual* (257) and Terman's *Measurement of Intelligence* (229). Before 1920 several other texts on measurement were published, and several numbers of the *Teachers College Contributions to Education* and the *Teachers College Record* described the construction and scaling of standard tests and scales. Since 1920 the most influential textual treatment of the present topic is probably Kelley's *Interpretation of Educational Measurements* (148), if one may select from a score or more of important treatments of measurements. Hull's *Aptitude Testing* (141) initiated another important movement.

Since 1920 various writers, particularly Monroe (175), McCall (165), and Ruch and Stoddard (207), have described in detail the steps in the construction of standard tests, together with the appropriate statistical procedures.

Validation Procedures

Criteria of validity—Monroe (175) discussed seven criteria for the validation of test items. Ruch and Stoddard (207) and Symonds (226) presented more extended statements of validation methods. Wood (260) presented evidence on the validity of the Thorndike Intelligence Examination and other Columbia tests. All such methods might be divided into two categories: "curricular" (analysis of courses of study, textbooks, examination questions, and the like; judgments of experts; analysis of errors; social utility; etc.) and "statistical" (correlations against independent criteria; increases in percents of successes in successive grades or ages, etc.).

Item validation—The fundamental method of selecting items (otherwise adjudged valid) in educational tests is that of determining the percent of successes in successive age or grade groups. Chapman (91) presented a variant procedure in trade testing where the subjects were classified as experts, journeymen, apprentices, and novices. In tests for use in single grades, as in many high-school subjects, some variation of the method of correlation of single items against total scores is used (91, 93, 155, 207),

Clark (93) presented a formula for the validity of test items which was attacked by Peatman (185) as yielding too small returns for labor involved. Whelden and Davies (256) divided the test group into three subgroups by relative scores on an outside criterion test as a method of selecting the most functional items. Lentz, Hirshstein, and Finch (155) compared four methods of selecting items, namely, percents of successes by highest and lowest thirds of group, Vincent's overlapping method (250), McCall's method (164), and Lentz's method of the summation of agreements (156). The highest- and lowest-third method proved best in general.

Smith (218) investigated the validity of judgments of difficulty of test items, as an initial stage in test construction, by experienced teachers, inexperienced teachers, and test experts. The judgments of experienced teachers proved most valid, with the test experts second, and the inexperienced teachers third; the average validity coefficients being .86, .80, and .76, respectively.

Validity as affected by specific determiners—Weidemann (252) demonstrated an important source of invalidity of test items through what he aptly calls *specific determiners*. He found that if the words *always* and *never* occurred in true-false items, such items were false two out of three times. Conversely, statements of degree or comparison were true in two-thirds of the cases. Brinkmeier and Ruch (87) analyzed 10,756 true-false items and found that 2,018 contained specific determiners and listed 15 categories of words or phrases which act to determine the truth or falsity of true-false items. Brinkmeier (88) showed sentence length to be a specific determiner; the longer the sentence, the greater likelihood of truth. Brinkmeier and Keys (86) described another such factor under the term of circumstantiality or general plausibility arising from wealth of detail.

Mathews' findings (162, 163) indicated that pupils select the left of two alternatives 33.8 percent oftener than the right, and that the upper is chosen 3.2 percent more frequently than the lower. Ruch and Meyer (196), however, failed to verify such biases.

Relative validity of different types of items—Workers in the field of new-type or objective examinations have attempted to discover whether completion, true-false, multiple-choice, and similar tests measure the same function. Brinkley (85) found essay and objective tests to be approximately equally valid. Ruch and DeGraff (199), assuming simple completion items to be valid, concluded that true-false and multiple-response tests measured the same functions as did the completion forms. Wood (261), using several different criteria, found no differences in the validities of different types of items. Eurich (115) reached similar conclusions; and the studies of Paterson and Langlie (183) and Ruch and Charles (198) are likewise in agreement. Remmers and others (193) found presumptive evidence, however, that certain individuals may do less well on one type of test than another.

Effects of instructions and scoring on validity—Several attempts have been made to determine the relative validity of scoring tests involving chance successes by the simple number right (R) method and by the correction formula, $S = \frac{R - W}{(n - 1)}$; where S is the score, R the number right, W the number wrong, and *n* the number of choices presented. (For two-response tests, for example, true-false, this formula reduces to S = R – W.) Ruch and Stoddard (197), Paterson and Langlie (183), Wood (261), Ruch and DeGraff (199), and others agree that the use of the formula lowers the reliability. Ben D. Wood (261), Ruch and DeGraff (199), and E. P. Wood (262) found, on the other hand, that the formula increases the validity while affecting the reliability adversely.

Ruch and DeGraff (199) also investigated the effects of instructions (a) to guess when in doubt and (b) to omit when in doubt. The latter seemed the more valid procedure, especially if the tests were scored by the formula for correction for chance.

Thurstone (238), Brinkley (85), Foster and Ruch (120), and Staffelbach (220) considered the ideal weightings to be applied to rights, wrongs, and omissions. These results are not altogether in agreement, possibly because of differences in the instructions. The evidence suggests that R – W scoring penalizes slightly for errors, and that omissions are significant in a theoretical score. Weidemann (253) also suggested the latter. Holzinger (135) published a proof that R and R – W scores correlate to unity when there are no omissions.

Time limits as a source of invalidity—A moot question has been whether time limits invalidate tests by making them measures of speed. May and Terman (264) found Army Alpha scores for single and double time correlated .965; Ruch and Koerth (205) obtained .966 in the same situation and .945 for regular and unlimited times; and Ruch (206) found even higher correspondences for regular and unlimited times for the Stanford Achievement Test and the Terman Group Test. These writers concluded that the timing of tests is no serious source of invalidity. Brigham (84) and Frank N. Freeman (122) drew exactly the opposite conclusion from these data, and held that the high correlations indicate that speed is the main variable. Frank S. Freeman (123) reported two sets of experiments where single and double time correlated .97 for speed and .78 for power. Longstaff and Porter (159) supported the former view. Peak and Boring (184) held that there is a marked correlation between speed in intelligence tests and speed in simple reactions. The final test should be whether the slow eventually catch up with the rapid. In no case was this found to be true.

Effects of group work on tests—The question whether working alone or in groups affects scores has been investigated by Weston and English (255)

and Farnsworth (116) with conflicting results. The latter found mean scores to be no higher in group testing than when individuals were tested alone.

Measures of Reliability of Tests

Concept of reliability—Spearman (219), Brown (90), and Kelley (148, 151, 152) developed the concept of test reliability and formulae for expressing errors in test scores. More general discussions were given by Symonds (225, 226), Ruch (204, 207), Mangold (160), and Foran (119). Symonds (225) listed twenty-five factors affecting reliability. Criticisms of reliability concepts and practices were brought forward by Crum (97), Lincoln (157), and Muenzinger (178).

Optimum reliability of test items—Otis in 1916 showed that the most reliable test was one on which the average score is 50 percent of the maximum score. Symonds (224) proved the same point and laid down six specific principles of optimum difficulty for maximum reliability. Cleeton (94) and T. Thurstone (241) discussed the same issue. Bliss (82, 83) objected to the use of percents of successes in selecting and arranging test items.

The Spearman-Brown Formula—Extended controversy has raged over the predictive accuracy of the Spearman-Brown Formula in estimating reliability of lengthened tests. Holzinger (134), Holzinger and Clayton (133), and Douglass and Cozens (105) reported over-prediction. On the contrary, Kelley (143), Gordon (127), Ruch, Ackerson, and Jackson (201), Wood (261), Remmers and others (191, 192), Lanier (154), Farnsworth (117), and Smith (217) found very close agreements of actual and predicted values in such different situations as discrimination of lifted weights, spelling tests, measures of musical talent, students' and teachers' judgments, etc. Slocombe (215, 216) and Thurstone (236) discussed the same formula critically. Shen (213, 214) derived and defended a formula for the standard error of predicted coefficients, which had been questioned by Holzinger and Clayton (133).

Reliability and sampling—Talbott and Ruch (228) showed the effects upon reliability of *intensive* and *extensive* sampling.

Statistical Treatments of Test Results

Types of norms—In 1920 the prevailing type of norm was the grade median or, more rarely, the grade mean. There has been a gradual shift toward age norms and variability units such as T-scores, sigma indexes, P. E. values, and percentiles.

Galton (124) and Woodworth (263) seem to have suggested the essential idea of comparable measures such as Kelley's standard measures (147, 152), McCall's T-scores (166), and Franzen's sigma indexes (121). Mon-

roe (174, 175) has argued for the use of age norms rather than grade averages. Numerous writers have discussed the advantages and limitations of different types of norms: Kelley (148), McCall (165), Monroe (175), Ruch and Stoddard (207), Willson (258), and Lindquist (158).

Scaling of tests—The early development of handwriting, composition, and language scales turned attention to the scaling of tests. McCall (165, 166), Van Wagenen (249), and Trabue (248) published typical procedures.

In a long series of papers, Thurstone (232, 234, 237, 239, 240) sought to present refinements of the earlier methods. His method of scaling sought to avoid the difficulties of the P. E. values which arise from unequal variabilities in successive age or grade groups. He re-scaled the Trabue language and Woody arithmetic scales by way of illustration. Holzinger (136) criticized certain of the assumptions basic to Thurstone's method and appeared to favor the calculation of scale values from a single group of wide age range. Courtis' isochron method (96) based upon the Gompertz equation also affords possibilities for scaling tests over wide age intervals.

Weighting of test items—In the past ten years the practice of weighting individual test elements has been subjected to general attack. Douglass and Spencer (104) found correlations of .975 to .999 for weighted and unweighted scores on four educational tests. As a result Douglass abandoned his weightings in revising his algebra tests. Holzinger (131) found unweighted scores to correlate .99 with weighted values by two common methods. Corroborative evidence was published by West (254), Corey (95), Scates and Noffsinger (212), Odell (180), Ruch and Meyer (196), Potthoff and Barnett (189), and others. It is probably better to abandon weightings in *tests* as contrasted with *scales* proper.

The A. Q. technic—Franzen's proposal (121) of the accomplishment quotient (A. Q.) in 1920 resulted in a flood of literature, at the outset favorable, but more recently largely antagonistic. It is to be noted that Monroe and Buckingham (173) and Pintner and Marshall (187) developed similar procedures independently. Difficulties in the A. Q. technic were soon pointed out by Toops and Symonds (247), Chapman (92), and Ruch (195), although Stebbins and Pechstein (222) defended it as a very valuable measure. Thomson and Pintner (230) demonstrated the danger of spurious index correlations in such quotients. Perhaps the most damaging evidence came from Symonds (223), Popenoe (188), and Odell (179), who showed the reliability of A. Q.'s to range from about .20 to .60 for existing mental and educational tests. Chapman (92), Herring (130), Huffaker (140), and others pointed out the necessarily large probable errors of quotients from unreliable measures. Foran (118) noted the typical finding that the reliabilities of such quotients are lower than the scores themselves. Wilson (259), Douglass and Huffaker (103), Morley (176), Rand (190), and McCrory (167) have all found the A. Q. procedure unsatisfactory. In the

meantime Kelley (148, 150) developed a technic for comparisons of mental and educational ratings which appears to rest upon a sound statistical basis. It seems at this date that the A. Q. procedure is more likely to be discarded than to be retained.

Time Saving and Cost Reduction Devices

Correlations by machines—Three devices for the calculation of correlations by machines have been reported, namely, the Hull (142), Dodd (102), and the Mendenhall-Warren-Hollerith methods (168, 251). The latter is an adaptation of the usual Hollerith sorting machine and is claimed to be the most rapid and economical device yet invented.

Plotting devices and correlation charts—Symonds (227) brought together as a convenient reference fifty-two variations of the Pearson Product-Moment Formula. Toops (246), Orleans (181), and Anderson and Toops (80) described simple apparatus for tabulating and computing correlations. Edgerton and Toops (113) simplified the calculation of intercorrelations by tables.

Convenient correlation charts or scatter diagram forms have been placed on the market by Holzinger (132), Kelley (145), Otis (182), Ruch and Stoddard (200), Ruger (208), Thurstone (233), and Toops (245). Dvorak (107, 108) has a chart for the computation of *eta* from grouped data and assumed means.

Tables, nomographs, abacs, and graphs—Holzinger prepared tables for the probable error of the correlation coefficient as found by the product-moment method (138) and, more recently, a set of tables for elementary statistical work (137). Edgerton and Paterson (114) computed a table for the sigma of a percentage. Cureton (100) published two tables to facilitate the computation of *rho*. Edgerton and Toops (113) constructed a table for determining increases of validity and reliability for lengthened tests up to $n = 15$. Edgerton (112) has a table for the probable error of R predicted by the Spearman-Brown Formula. Masters and Upshall (161) tabled the probable errors of certain inter-percentile ranges.

Dunlap and Kurtz (106) published a valuable handbook for statisticians in which appear twenty-eight nomographs, twelve tables, and a collection of the more useful formulae.

Nomographs, abacs, or graphs have been prepared to facilitate computations with common formulae by Cureton and Dunlap (98, 99), Griffin (128, 129), Edgerton (109, 111), Toops and Edgerton (244), and Rulon (210).

Partial and multiple correlation methods—The increasing use of partial and multiple correlation methods in educational and mental analysis and prediction has encouraged the search for more economical methods of handling large numbers of variables. The method of Yule has been further

simplified, in turn, by Rosenow (194), Kelley (144, 152, 153), Huffaker (139), and Bathurst (81).

Kelley and Salisbury (149, 211) presented an iteration method for successive approximation of regression weights for large numbers of variables. Tolley and Ezekiel (242, 243) claimed that the older Doolittle method is more economical. Kelley and McNemar (146) made the counter-assertion that the iteration method is more economical for ten or more variables. Garrett (125) further simplified the Doolittle method and Peters and Wykes (186) prepared work sheets for this method.

Critical Issues

Critical issues looking toward the improvement of educational measurement might be discussed at almost any length. In view of space limitations, the reviewer will avail himself again of the opportunity to mention the very great significance of Kelley's *Interpretation of Educational Measurements* (148). Most of the essential issues and "next steps" are ably presented therein. The following further comments may supplement Kelley's discussions.

1. There are in use today at least one thousand different educational and mental tests. Convincing critical and statistical data on the validity, reliability, and norms of these measures are available in probably less than 10 percent of the cases. The publication of such crucial information is an ethical obligation of the test author and publisher. Ruch (202) suggested the minimal requirements in such reporting.

2. In view of the situation just mentioned, there is an urgent need for comparative studies of the relative values of existing tests. In most subjects, this need is probably more insistent than the production of new tests. Kelley and a group of experts partially met the situation (148: 214-348). Gates (126), Monroe (170), Ruch and his students (79, 101, 196, 203, 207), Mosher (177), and Broom and others (89) made scattered, individual attempts to evaluate tests of reading, history, geography, physics, arithmetic, and other subjects of study. The gross unreliability of many published norms was strikingly demonstrated by Adams (79), who found that 8 of the best known arithmetic tests rated the mean performance of 152 pupils all the way from fifth-grade to eleventh-grade achievement, depending upon which test was employed.

3. The reliabilities of all but a few existing tests are far too low for the measurement of individuals, as contrasted with evaluation of groups. Monroe and others (172) found an average reliability of .67 for 21 standard tests. Ruch (204: 142-4) computed 149 such reliability coefficients and found the central value to be .69. In view of Kelley's standards of required reliabilities (148: 28-9), 131 of the 149 tests are serviceable only for group measurement; about 10 are adequate as measures of individuals; and only 5 or 6 will justify attempts to compare differences in achievement in different school subjects (A. Q.'s or more valid technics). These facts should dispose of the demand for shorter tests; longer and more reliable ones are indicated.

4. Mislabelling of tests is the rule rather than the exception in such titles as *diagnostic* and *prognostic*. Very few diagnostic tests show sufficient reliability of *total* scores for accurate measurement, not to mention the unreliability of the sub-tests individually. Few prognosis tests predict better than a correlation of .60 and frequently subjectmatter and intelligence test data are already at hand which would enable equally as good or a better prediction, if properly weighted and combined. The reviewer has considerable unpublished evidence in support of this contention.

5. There is urgent need for a fact-finding organization which will undertake impartial, experimental, and statistical evaluations of tests—validity, reliability, legitimate uses, accuracy of norms, and the like. This might lead to the listing of satisfactory tests in the various subjectmatter divisions in much the same way that Consumers' Research, Inc. is attempting to furnish reliable information to the average buyer. The reviewer has indeed attempted to initiate such a fact-finding project, but without success to date. Independent workers in this field are few as yet, the task is tremendous, and to leave such determinations to authors of tests and publishers is likely only to continue the present chaotic conditions.

CHAPTER IV

Recent Developments in Testing for Guidance

ITEMIZING the findings of published articles in a given field of research may fail to indicate in proper perspective the larger problems that lie behind the bits of evidence which have found their way into print. In order to be able to present a broad view of what is being attempted in guidance, personal letters on the subject were solicited from approximately fifty American psychologists and educators who were known to be actively interested in guidance. It will be understood, therefore, that points of view and quotations ascribed in this summary to certain workers without the citation of specific reference numbers were taken directly from these personal letters to the reviewer.

Testing for Selection

The distinction between the guidance of an individual and the selection of an employee or student for a particular type of work is gradually being recognized. The employment managers of factories and perhaps the admission officers of many of our schools, are interested in selecting only those persons who can do the work successfully. This point of view, as Harry D. Kitson points out, is not that of guidance:

> In some occupations, with reference to certain specific jobs, some psychological tests have been able to select good workers with a satisfactory degree of accuracy. But this is not vocational guidance. It is vocational selection. To do vocational guidance we should try to help the unsuccessful applicants to find satisfactory vocations.

Herbert A. Toops states that "the twelve thousand or so freshmen of Ohio colleges could be recruited twice over from high-school graduates who have at least a fifty-fifty chance of graduating." It is probable that almost any commercial or industrial organization could, especially during the present period of economic depression, replace satisfactorily its entire staff from the ranks of unemployed persons by using available tests in their selection. Stanton (298: 43), in selecting students with tested musical ability for musical training, argued that "competent teachers are deserving of the best talent we can give them. Why should a school obtain the best of teachers and give them any pupils who wish to study, poor talent as well as good talent?" Tests and other means of selection for many types of work are now available, although relatively little has yet been done with them by most schools and industries.

One good illustration of effective research with regard to selective devices is the work with placement tests that has been going forward under Stoddard (300, 301) and Seashore (294, 296) at the University of Iowa. Miller (283: 112) reported that "reading comprehension tests which demand the under-

standing of the application of principles and which are based upon the particular subject to be taken constitute the best single measure of aptitude employed in these examinations." Viteles (310: 200-322) presented an excellent summary of much of the research that has been done on the selection of workers in industry and business.

In all of this research on selection, it is becoming increasingly evident that objective measures should be substituted where possible for the less reliable criteria that have commonly been employed. As Richard D. Allen says, "Every day brings evidences of the danger of guess work in cases where measurement is possible." Even verified employment records, showing the number of years an applicant has worked at a given job, are surprisingly lacking in their validity as measures of his skill or ability to do the work. In examining the academic abilities of 282 unemployed persons in Minneapolis who had graduated from high school but taken no additional training, more than 1 percent of them were found to possess fifth-grade ability, 2 percent sixth-grade ability, 3 percent seventh-grade ability, 6 percent eighth-grade ability, and so on. Equal lack of real validity was found in occupational experiences of men who had worked from ten to twenty years in the same occupation. Men who have earned their living for years as machine-tool operatives sometimes possess less actual skill and less actual knowledge of their tasks than youngsters who have had only a few days of experience.

Commenting upon officials who fail to recognize this lack of equality in ability among persons who have had equal opportunities to acquire it, Remmers (293: 28) remarked that they "are concerned with what goes into the process and not with what comes out." A similar note appears in Link's (282) comment that "the cultural value of education resides not so much in the courses chosen, as in what these courses do for the individual." Industry and education are gradually learning that it is much safer to rely upon objective measures of present ability than upon mere records of time served at a certain type of work.

The Scope of Guidance

It is possible that millions of workers are earning a living at tasks to which they are less well adapted than they would be to certain other tasks. Most of these persons, poorly adjusted from the point of view of individual success and happiness, are probably fairly efficient from the point of view of business and industry. They may be well selected, but they have not been well guided. In guidance, as Viteles (311: 339) remarked, "the point of orientation must always be the individual and his future."

It will never be possible in a rapidly changing world to have everyone working at the particular task which he can do best, but that fact does not excuse us from making serious efforts to approach such an ideal. Any organization of society which fails to contribute to the personal adjustment

and happiness of its citizens will gradually disintegrate, but every contribution made to the personal satisfaction and contentment of its individual members will add strength and life to the organization which makes the contribution possible. Guidance aims, therefore, not only to help the individual to find his appropriate place in society, but also to strengthen society itself by developing greater satisfaction and loyalty in each of its members.

In the practice of guidance one must consider one's work incomplete until the most appropriate adjustments possible have been indicated for each individual. Ignorance of some important factor in the nature of the individual cannot release one from one's responsibility for making an incorrect diagnosis. It is obvious that no adviser can ever know all about every trait possessed by those individuals to whom he attempts to give guidance, but, here again, the inability to attain perfection should not prevent us from attempting to approach it as closely as possible. Every aid that science can give to an individual in seeing himself and his own characteristics more clearly, and in understanding more fully the place in which his talents would be of greatest value to himself and to society, is certainly worth the effort involved in providing it.

From this point of view, guidance and education are very intimately related. As a matter of fact, educational guidance, social guidance, emotional guidance, vocational guidance, and all other desirable types of guidance are merely different phases of a single program whose purpose is to build the happiest and most fully integrated personality possible upon the foundation with which nature and previous experience have provided the individual. The principles of guidance are the same in all fields. While occupational guidance is most often discussed, it is only one phase of the total process, and it should not be viewed as an independent task. Occupational guidance may be used, however, to illustrate the problems and procedures that characterize the entire field.

Success and personal satisfaction in a given type of work involves the presence of a distinctive combination of abilities, interests, preferences, and other personal characteristics. While the possession of a few special traits may be indispensable in the occupation, there are other traits which add materially to the general excellence of the individual's adjustment. Definite knowledge of the distinctive patterns of traits which characterize the successful workers in each occupation is essential in one who would give occupational guidance.

The systematic determination of these patterns in an objective manner has made relatively little progress. The work of the Employment Stabilization Research Institute of the University of Minnesota (280, 288, 308) has indicated one of the methods by which such determinations might be made. Strong (302, 303) at Stanford University and other workers (268, 269, 306) have been determining the interest patterns of successful persons in

various professions, but there is just as great a need for reliably determined patterns of ability, patterns of physical condition, patterns of attitude, and patterns of social behavior. Until these distinctive patterns are made available, much of the vocational guidance offered will continue to be in the nature of "the blind leading the blind."

In addition to knowing the distinctive patterns of characteristics possessed by successful and well-adjusted workers in various occupations, the person who attempts to give vocational guidance should also know how these well-adjusted workers came to have these patterns. How early in life do various features of these patterns appear? To what extent are the characteristic traits of successful workers developed by training and experience, and to what extent are they the result of original nature? The answers to these questions must be sought by means of cumulative records for individuals. Objective records of test results during the elementary-school period, during the high-school period, and during late adolescence must be studied in relation to equally reliable records of the training and experience received at different periods by these same individuals. Analyses of such cumulative records will show how early in life mechanical abilities, clerical aptitudes, artistic appreciations, submissive personalities, and selling interests crystallize sufficiently to be truly indicative of their ultimate development. At the present time we have very little reliable information regarding these critical issues.

This discussion calls attention rather sharply to the fact that very little evidence has yet been found to indicate that educational tests have any great value in vocational guidance. It is true, of course, that one's scores in objective educational tests have recognized usefulness in predicting one's probable success in school and college courses, but these values are only indirectly concerned with vocational guidance. Consistently low scores in educational tests would generally be accepted as indicative of probable failure in law, engineering, and other professions; but very little is yet known regarding the specific patterns of educational test scores characteristic of those persons who later become successful in various occupations.

Here again, a broad and fruitful field of research awaits the investigators who can identify a sufficient number of successful, well-adjusted workers in different occupations and discover from their cumulative school records the test-score patterns which characterized them in early life. If the records are sufficiently full to show the specific educational and occupational experiences which these persons met between the time they took the tests and the time they achieved success and happiness in their occupations, significant additions will be made to our knowledge of desirable educational procedures as well as to our technics of guidance.

Forward Steps in Guidance

The demonstration in the World War of the usefulness of so-called "general intelligence tests" (290, 315) led to extensive experimentation with

mental tests in vocational and educational selection. The limitations of such measures are beginning to be determined, and the tendency at present is to use in their stead measures of specific types of ability. Stoddard (301: 23) reported that "a number of placement examinations lead to a profile of one's mental-educational skills, which in the case of adults is more intelligible and more significant than a single measure, such as I.Q.," and (301 : 92) that "partial and multiple correlations demonstrated the superiority of placement examinations over high-school achievement and the traditional intelligence test as a device for predicting college success in a subject." Comparable results have been obtained by other investigators of educational selection (276, 293, 298) and by practically all those who have investigated problems of vocational selection (273, 308, 311).

The large number of different traits that characterize successful workers in each different occupation has made it evident that vocational guidance cannot confidently be offered to an individual without a very wide sampling of his characteristics. The Employment Stabilization Research Institute of the University of Minnesota (280, 288) in their examinations of approximately four thousand unemployed persons used a uniform program of individual tests and examining schedules requiring of each individual approximately six hours, supplemented by such special examinations and follow-up studies as were found necessary. The Minnesota examinations included personal, social, educational, and occupational histories, carefully checked by well-trained industrial-social case workers and cleared through the confidential exchanges of the community social agencies; complete health and physical examinations, supplemented by routine fluroscopic and bio-chemical tests; measures of physical strength, visual and auditory acuity, color blindness, and the like; tests of academic achievement, clerical aptitudes, mechanical abilities, dexterity in the use of hands, fingers, and small tools; and measures of interest patterns, likes, dislikes, and personality traits. Trade tests of skill, knowledge, and appreciation were also used freely.

Link (282) reported that Hanna found "from five to fifteen hours time essential to an adequate vocational analysis," and that Viteles found "five hours the minimum time necessary in the simpler cases." Viteles himself (311: 335) stated that "underlying the work of this guidance clinic is the point of view that adequate guidance involves a consideration of all the psychological, social, economic, and physical factors which may affect the progress of an individual in a vocation."

Interpretation of all these data regarding an individual, even if one were adequately supplied with complete information regarding the distinctive patterns of traits characteristic of each possible occupation, would be a task calling for the utmost of mature judgment and sagacity. Heller (272:437) in reporting on the best practice of industrial psychology in Switzerland, said, "The psychological diagnosis attempts to survey the total personality

of the individual." Viteles (311:338) declared that "the problem of guidance is that of weighing every element in the situation—of establishing a balance between the highly complex interrelated variables which influence individual adjustment." The seriousness of this responsibility leads R. S. Uhrbrock to ask, "By what right does a vocational counselor, for the most part unversed in clinical methods or in statistical technics, administer tests and then undertake to interpret the results and give vocational advice?" Brotemarkle (267:258) also believed that "psychometricians are not adequately trained to give the analytical diagnosis basic to the solution of individual problems."

Definite progress is being made in determining the statistical reliability and validity of various specific measures. There would be a certain amount of advantage scientifically if those who devise tests could register them and have them assigned certain numbers rather than names. O'Connor (286), for example, preferred to discuss "Work-sample No. 16" rather than the "Finger Dexterity Test." Any name assigned to a test leads those who hear it to expect from the test other types of information than it can possibly provide. The painstaking use of correlations (273, 287, 300), tetrad differences (297), path coefficients (314), and iteration methods (277) in teasing out the real meanings of test scores will ultimately be of great value in guidance, since it will make clear to us the degree to which different tests are measuring the same traits.

The development of tests for other traits than achievements and abilities is making relatively slow progress. Hartshorne and May (271) made significant progress in measuring social and ethical behavior, and Thurstone (307) did valuable work in measuring attitudes. Strong's (303) testing of vocational interests was a contribution of marked importance. Mary H. S. Hayes, for example, writes that in her judgment "Strong's interest tests come nearest to being the most significant development, in spite of the fact that they are not applicable to the younger group."

There is great need, however, for more objective tests of determination to succeed, willingness to sacrifice immediate comfort for ultimate success, physical attractiveness, and other traits that are not closely related to mere abilities. Daniel Starch writes, for example, that "achievement in occupations depends to a larger extent upon such qualities as initiative, aggressiveness, and industry than is commonly realized. Tests to date have not apparently measured these qualities with sufficient reliability."

Other Critical Issues

One of the difficult tasks which is slowing up the progress of testing for guidance is that of identifying well-adjusted individuals in each occupation. The so-called "democratic organization" of society in America tends to disturb even those persons who are quite happy in their work, since some

of them are subject constantly to the temptation of wishing for social prestige, greater political power, or larger financial return, even though they may not be at all typical in their trait patterns of the persons who are well adjusted in these other positions which they covet. In setting up standards, we must be very careful to select as representative of each occupation, therefore, only those who are clearly adapted to their positions.

Another difficulty lies in the lack of adequate organization for obtaining the patterns of traits typical of well-adjusted persons in the various walks of life. An individual research worker can do relatively little by himself, and what he does is in danger of being only local in its significance. A national clearing-house and research center with adequate funds for the nation-wide determination of occupational patterns, each with its local variants, is very greatly needed. Such an organization would also be of tremendous public service, in revising our scheme of occupational classification, which years of social and technological changes have rendered quite obsolete.

One of the problems on which further experimentation is necessary concerns the methods one may use to represent the patterns or combinations of traits in an individual or in an occupational group. A graphic profile has been used (308), and Hull (273) has proposed a differential index to be derived statistically by multiple correlation technics.

Whatever the method used in recognizing the individual's pattern of traits, psychological insight into the individual's whole personality, and a full understanding of the working conditions in the various occupations open to him, should be possessed by his occupational adviser. This means, among other things, that the vocational adviser must be selected by means of the most rigid selective devices available and then trained, not only in the clinical methods of vocational psychology, but also in first-hand contacts with the actual conditions in business and industry. As R. S. Uhrbrock says, "Any movements that provide first-hand vocational experience for counselors are worth fostering."

Still another problem that must be attacked cautiously is that of developing public confidence in the value of information obtained by these technics. It is easy to claim too much validity for a diagnosis; yet millions of dollars are being thrown away annually by persons who are consulting palm readers, phrenologists, graphologists, and physiognomists in serious efforts to obtain better adjustments to their occupations. Objective tests, if properly interpreted in the light of scientific studies, provide the soundest basis for really valid guidance, but the results of two or three tests on a given individual should not ordinarily be considered adequate. There is no short-cut to a valid occupational diagnosis.

Vocational guidance should never become vocational control, and yet some effective way must be discovered to persuade unadjusted persons to consider seriously the results of suggestions growing out of careful occu-

pational diagnoses. Traditions and beliefs regarding the greater social dignity of certain occupations are hard to overcome. Harry J. Baker says, "It seems to me that one of the most crucial problems in guidance is ways and means of changing and molding the vocational aspirations of dull and average individuals into lines of activity within their abilities and to do this without undue shock and with a positive rather than a negative psychological approach."

The reviewer has felt for many years that one of the first steps in breaking down popular prejudices against certain kinds of work is to abandon the use, in discussing workers and occupations, of such adjectives as dull, inferior, superior, intelligent, and the like. From a social point of view the garbage collector who likes his work, and who has all the appropriate characteristics for that particular job, is a much better citizen than the man who, with the characteristic trait patterns of a mule driver, is trying to manage an employment office or banking business. Perhaps it is impossible in a capitalistic economic system to substitute the concept of adequacy of personal adjustment for that of adequacy of financial standing as a basis for social prestige, but we should certainly do everything possible to remove artificial stigmas of every kind from all types of useful service.

CHAPTER V

Recent Developments in the Uses of Tests

SCOPE OF TESTING MOVEMENT

THE scope of measurement in education is probably not yet generally appreciated. When one considers the kind and number of research studies in education since the advent of the testing movement in contrast with those of the previous two or ten decades, the realization of the potency of this new instrument is inescapable. Testing procedures are now a matter of course in the attack on educational problems everywhere. Twenty years ago tests were novelties—technics of investigation consisted largely of the compilation of opinions. Today the use of educational tests has become almost as commonplace as that of textbooks. In the more progressive school systems, teachers utilize various forms of educational tests continuously. Thousands of students of education are making use of this relatively new device. The ultimate purpose may, in general, be said to be the improvement of instruction. This is often sought indirectly through changes in administration, in organization of schools, in classification of pupils, in educational and vocational guidance, and so on; but as a rule some form of measurement constitutes the basis. Standardized and unstandardized educational tests have thus, in large measure, become everyday working tools for teacher, principal, and superintendent. Many investigations, of course, are never published, but a considerable number of bibliographies have already been compiled. A list published by Monroe and his associates (396) in 1928 contained 3,650 references to educational researches. A large number of these involved the use of tests. The *Education Index* for January, 1929, to June, 1932, (349) alone listed 139 articles under "Tests and Scales," 7 under "Achievement Quotients," 36 under "Achievement Tests," and 55 under "Educational Measurements." Many of these references are themselves compilations of articles on the same topic. The United States Office of Education has published annually since 1926-27 a *Bibliography of Research Studies in Education* (451, 452, 453, 454) which includes hundreds of studies utilizing tests as the essential instruments of research. Two hundred and fourteen studies out of a total of 4,651 for 1929-30 were classified under "Testing and Research" (454). A large number of other studies listed in this single bibliography made use of tests. The University of Illinois has also published an index of indexes (394), which includes a large number of studies based on educational tests, and an annotated bibliography of graduate theses in education (364). The *Teachers College Record* for January, 1932, (464) contained a bibliography on sources useful in determining research completed or under way, including the studies of the National Education Association; the United States Office of Education;

the sources for locating research undertaken by individual institutions, such as the larger universities; theses and dissertations; and the abstracts found in the past issues of the *Review of Educational Research*. Such summaries of research in special subjects as those prepared by Gray (361, 362), Buswell (331, 332), Monroe and Engelhart (392), and Lyman (384) likewise presented annotated bibliographies describing research studies in which large use was made of educational tests.

This wide range of references is evidence of the growth of the testing movement throughout the United States in little more than a decade. So vast a literature cannot be reviewed in the brief space here available. Under the circumstances it seems a more practical procedure to cite typical contributions without any attempt to be exhaustive. Probably as much as a "five foot shelf" of texts specifically addressed to the uses of tests has appeared since the beginning of the test movement. Among the chief recent contributors may be mentioned Monroe (395), McCall (388), Trabue (445), Courtis (341), Gilliland, Jordan, and Freeman (359), Pressey and Pressey (414), Ruch and Stoddard (425), Wilson and Hoke (463), Smith and Wright (430), Orleans and Sealy (407), Odell (406), Greene and Jorgensen (363), Ruch (423), Kelley (378), Hull (372), Van Wagenen (455), Hildreth (368), Madsen (386), Russell (426), Michell (390), and Tiegs and Crawford (443).

Generally speaking, these authors have addressed themselves to the use of tests. Achievement tests have received considerable attention. The *Measurement and Adjustment Series* of texts edited by Terman, which includes some of the authors already mentioned and in addition such writers as Dickson, Fenton, Goodenough, Otis, Stidham, Wells, and Wood, deals with the general problem of pupil testing and adjustment. A closely related series of statistical texts also dealing especially with educational measurements and their interpretations has appeared. Among these may be mentioned books by Otis (411), Holzinger (371), Garrett (354), Lincoln (383), Thurstone (442), Macdonald (385), Kelley (378), Dunlap and Kurtz (346), Odell (406), and others. All this literature has appeared subsequent to such basic studies as those by Galton (353), Cattell (334), Thorndike (441), and Terman (438, 439, 440).

TYPES OF USES OF TESTS

A mere listing of ways in which educational tests affect educational theory and practice serves to emphasize the recent wide influence of this relatively new technic. Among such uses may be mentioned the following:

(1) *Determining and evaluating administrative policies*, including the classification of pupils, provision for individual differences, standardization of teachers' marks, curriculum construction, and supervisory activities.

(2) *Setting up objectives and evaluating the products of the educational program.*

(3) *Evaluating methods of teaching.*

(4) *Improving learning* through a discovery of learning difficulty, the sources of motivation, and the uses of self-teaching test materials.

1. Determining and Evaluating Administrative Policies

Educational tests, as a measure of the effectiveness of administrative policies, got under way in the surveys made either by local school boards, by special commissions, or by bureaus of research. In the 194 city school surveys made since 1910 and reported by Caswell (333), more than half involve the use of tests as basic data for evaluating the school system. In the Baltimore survey of 1921, for example, the status of pupils revealed by standardized arithmetic tests called attention to the fact that too much time was being given to this subject and too little to other subjects. The recommendations of the surveyors based on test data led to a change in the administrative policy in this respect. Local surveys, such as the semi-annual instructional survey conducted by the Baltimore Bureau of Research (433), have influenced practically every phase of the school system. This is likewise true in Providence, Philadelphia, Cleveland, Detroit, and other cities.

Classification of pupils—It was inevitable that the possibility of more accurate measurement of the educational product of the classroom should lead to frequent uses of educational aptitude tests in the sorting and resorting of pupils within schools and within classes. While intelligence tests have from the beginning served as the more common means for so-called homogeneous grouping, the many controversial issues involved have served to deflect more and more interest toward achievement tests whose significance is at least thought by most persons to be less uncertain. Grouping pupils in and within classes on the basis of achievement scores in successive subjects has become a typical use of educational test data. The reports of 555 superintendents to the National Education Association (402) indicate how frequently this practice is followed. A detailed statement as to how three cities use this procedure is shown by Baltimore, Colorado Springs, and New York City. Chism (335) reported that 67.5 percent of the elementary schools in 490 cities with a population of 2,500 to 100,000 use standardized educational tests as a basis of classification.

As this practice developed, questions arose as to the relative merits of several methods of classification, and again educational tests became effective means of aiding the evaluation. Burr's examination (329) of the educational achievements of homogeneous groups with emphasis on variability as a supplement to central tendency, concluded that there was great overlapping of achievements of groups as sectioned in the six cities studied. Purdom (416) found that first year high-school pupils do not gain more in English and algebra than pupils in heterogeneous sections when the results are measured by standardized tests. Keliher's investigation (376) called attention to the effects of homogeneous grouping not measured by educational tests. Hollingshead (369) evaluated the use of certain educational tests and mental measurements for purposes of classification.

As the problems of classification loom up, prognosis becomes an end for using tests. Courtis (341) studied how reliably the success of a child can

be predicted from the measurement of a few basic factors with existing tests and scales. He reported that boys within the age range and school conditions studied have been proved to succeed in their school work in differing degrees primarily because of differences in the maturity or development factor best represented by age. The comment is also noteworthy that approximately 7 percent of the present errors of prediction of Stanford scores have been proved to be caused by imperfections in the measuring instruments themselves. In checking the relative values of individualized versus group instruction when the pupils under both plans go on to higher schools, Washburne, Vogel, and Gray (459) found the results indicate that the mastery of the fundamental facts in arithmetic, reading, and language as measured by standardized tests is facilitated somewhat for most pupils by the Winnetka technic. One classification problem, namely, class size, has been thoroughly investigated by Smith (429). Her contribution is significant, not only in its resulting information concerning the conditions under which large classes can be effectively handled, but also in her skillful use of tests. In addition to the series of tests incident to the pairing of the groups, a program of achievement testing was carried on throughout the year. Pre-tests were used in many units of work, and quarterly tests in the common English skills. Where standardized tests were available, they were used in alternating forms at the beginning and at the end of the year and, in some cases, at the end of each quarter. Where standardized tests were not available, objective tests based upon the particular content of the course were devised.

Provision for individual differences—Growing out of classification studies comes the realization of wide-spread individual differences. In the procedures used for providing for these differences, educational tests function, not only to discover differences and to evaluate procedures, but also as an integral part of the teaching technics. The National Society for the Study of Education (403:xii), for example, called attention to the fact that complete diagnostic tests need to be prepared on each unit of achievement, and that self-instructive and self-corrective practice materials are to be prepared to enable children to get ready for the tests individually or to repair deficiencies shown by the tests. Buckingham (327) set forth a program of individualized instruction on the basis of testing. Washburne (456) supported his philosophy of individualized instruction with test data showing that high-school pupils from Winnetka's individualized schools when compared with pupils from other localities are above the average of all other pupils in four classes. Using educational tests as one basis in equating groups and as a measure of growth, Broening (325) made an experimentally determined study, which revealed that individualized technics based on initial test data and self-corrective test materials brought about greater achievement among junior high-school pupils in geography. Elective courses and preventive classes utilizing educational test data, either to select individuals or to measure achievement, have also been developed to provide for individual differences (319).

Remedial teaching—An essential step in diagnostic or remedial teaching is a skillful use of educational tests. Sangren (427) gave a thorough-going report on the use of tests in the improvement of reading. Gates (355, 356, 357, 358) gave a detailed account of a system of measuring achievements, diagnosing difficulties, and conducting test-determined instruction in reading. Objective test data are presented indicating the value of the procedures set up. Similar work was conducted by Courtis and Clapp in arithmetic. Metcalf (389) emphasized the need for using scientific testing in the analysis and classification of pupils in an effort at correct educational guidance. Dvorak and English (347), in measuring the efficiency of remedial teaching based on the Stanford Achievement Tests, found that their program of test-determined teaching produced from .99 to 2.3 years more than the expected or regular .5 year growth; and claimed that the reason the pupils in their school had been so retarded was that though the teachers knew in a *general way* that the pupils were handicapped, they lacked the scientific and quantitative information which only the standardized survey and diagnostic tests can give. Monroe (391) made use of tests in developing methods of diagnosis and treatment of cases of reading disability.

Promotion of pupils—Ample evidence shows that the technic of advancing pupils to a higher grade is still far from objective. One of the yet unsolved problems is the weight which should be attached to the results of objective tests where available. That achievement tests, however, should play a part in determining promotion seems obvious, but there has been, apparently, little progress as yet in the field of test-determined standards of promotion. The beginning made in Baltimore, reported by Kramer (381), Douglass (345), and Frazee (352), indicates how, as a result of the city-wide programs of testing in arithmetic and reading, standards of attainment have been set up for the skill subjects in the elementary grades and in some subjects on the secondary level. These test data assist teachers in determining which pupils have made sufficient achievement in the skills indicated to warrant their promotion at least in the subject concerned. By allowing a deviation below the standards by as much as a half grade, the danger of injustice is removed, and at the same time the evils of courtesy promotions and low standards on the part of individual teachers are eliminated. The plan has further administrative advantages in that it fixes responsibility directly on principal, supervisor, and teacher for the scholastic standing of their school. For those pupils to whom the regular offerings are unsuited, a special program suited to their ages and abilities is indicated, and the administration of this program is greatly facilitated by the application of objective standards. Although the operation of the system of standards here described is far from automatic and objective, it is well removed from the highly subjective conditions which obtained twenty years ago.

The *Ninth Yearbook* of the Department of Superintendence offers (402: 56-64) some excellent suggestions concerning means by which the classroom teacher may reduce failure. The three items of greatest frequency directly concern the use of tests:

1. Use achievement and diagnostic tests followed up by special help and remedial work—test for deficiencies and diagnose pupil difficulties in each subject.
2. Give individual attention to pupil needs and interests. This directly implies some objective measuring instrument to determine needs.
3. Group according to ability, differentiate courses of study, and apply teaching methods suitable to each ability.

In response to the inquiry, "What means do supervisors find most successful in bringing about a wider application of acceptable principles relative to pupil promotion?", fifty-five selected supervisors who were asked to contribute to the *Ninth Yearbook* made most frequent reference to the use of standardized tests to supplement teacher judgment. Collier and Miller (337) and Tyndall (449) also showed the use of achievement tests in the solution of promotion problems.

Standardization of teachers' marks—Ruch (423) proposed the use of standardized achievement tests to stabilize the marking system of a school organization. If the normal curve is used in marking, "in the first place we should disabuse ourselves of the idea that the normal curve (or any other mathematical concept) will tell us exactly how many pupils should receive A, B, C, etc." However, in systems using five letters, the approximate distribution, based upon "the assumption of chance distribution of pupil abilities" is: A's, 6 percent; B's, 25 percent; C's, 38 percent; D's, 25 percent; and E's, 6 percent. Such a distribution would hold for large numbers of pupils, and teachers are justified in objecting to its mechanical application to small classes. Marked departures, however, should be based upon demonstrated variation from normal conditions. Over a long period, marks should approximate the normal distribution and marked variations may be questioned. Ruch makes a number of significant proposals in marking, among which is this statement: "Give a standard test as a check on grades given." Taylor's study (437) shows that where objective teacher-made tests are used, the teacher's marks are more reliable than when only essay tests are used.

Supervisory activities—Present-day supervision in the better school systems rests largely on test-determined achievement. The tools of measurement are continually being utilized as one of the most effective means of supervision. Barr (318) brought to light the need of measuring devices which will harmonize with the tenets of the new education and gave a helpful discussion of the use of test data in a supervisory program. While many inferences drawn from test data by supervisors are probably unscientific, it may be said that a definite beginning has been made in this technic. In the Baltimore school system, for example, copies of all test results (ob-

tained regularly at the beginning of each term) are furnished to the supervisory force. The achievement of each class and of each pupil is scrutinized by the supervisor before making visits to the school, and recommendations for improvement of instruction are based largely on the test results (420). Cutright (344) showed how educational tests were used in developing scientific instructional material, in the determination of an effective teaching method, and in equating groups of pupils used in each experiment. Burton (330) reported the Detroit study of 1918, in which supervision was evaluated by measuring children's achievement through standard tests, Crabbs's 1925 investigation in measuring efficiency in supervision and teaching through the use of educational tests, and Brueckner's study of work type reading—a test-determined supervisory program. Seaton and Pressey (428) indicated how a report of the results of the tests given in October, 1931, was prepared in order that teachers and principals may have a guide to aid them in improving the work in composition in their schools. Knight (379) showed the uses of tests in surveying instruction. Coy (342) reported a study of the use of the accomplishment quotient in grades 3A to 6A as a measure of teaching efficiency. Michell (390: 169-175) brought out the teaching values in new-type history tests. OBrien (405) indicated how tests were made a part of a supervisory program in geography and history. Holroyd (370) discussed a supervisory project in educational measurement.

Curriculum construction—Perhaps the most interesting recently developed use of tests is in connection with curriculum research. This type of investigation reaches into the selection of subjectmatter, grade placement of units of experience, and methods of organizing the curricular offerings. The *Sixth Yearbook* of the Department of Superintendence (401:325) recognized the place of educational tests in curriculum research. Burch (328) utilized silent reading comprehension tests in the determination of the content in literature suitable for junior and senior high-school students. Collings (338) used standardized educational tests as part of his basis for equating groups. Standardized tests were also used as a vital part of the measurement of outcomes in the experimental and control groups, proving the advantages of the project curriculum. Guiler (365) revealed important data which are being used by curriculum workers when he analyzed the results of the O'Rourke Test given to 240,000 pupils. Harap (367) showed a need for formulating tests and practice material as part of curriculum programs. Irion (375) indicated a successful attack on curriculum problems when he analyzed literary comprehension into five elements, devising tests of each element for four different types of literature, using intelligence and standardized reading tests for comparison. Washburne (456) reported a five year study on adjusting the arithmetic curriculum to the child, "foundation arithmetic tests" being used in the initial and final measurements which determine allocation of arithmetic topics. He also described the use

of standardized reading tests in determining the difficulty of the reading material to be given the individual child at any particular stage of his development. Wood and Freeman (465) used educational and intelligence tests in determining the influences of the typewriter in the elementary-school classroom. Adams (316) used educational tests as a background for trying out the geography and history in the intermediate grades. Broening (323) used cumulative test data and initial and final test scores in the cooperative English research studies undertaken in Baltimore. Rankin (418) presented survey technics for the experimental determination of the value of materials and methods. Rolker (421) studied the spread of ability in arithmetic and its relation to standards of promotion and course of study revision. Bamesberger (317) compared outcomes of two types of social science courses of study through controlled investigations in a limited situation. Standardized reading tests were used to equate the two groups studied, the outcomes being measured by carefully constructed objective tests on the sections of the subjectmatter in question.

2. Setting Up Objectives and Evaluating the Products of the Educational Program

Eells (350) found in his study of the tests used in seventy-two published school surveys that arithmetic and reading are the subjects in which tests are most often used. General intelligence, spelling, and penmanship are next in frequency, and no other subject is tested in half the surveys. Hence, over a period of years the availability of standardized tests controlled to a large extent what objectives of education were objectively measured. A case in point is the development of the technics of silent reading as a major goal of reading instruction. As educators became more critical of the direct relationship between what is measured in education and what receives emphasis in teaching, steps were taken (1) to *prepare standardized tests covering more of the objectives set up,* and (2) *to utilize the objective test technic in the measurement of the entire list of acceptable objectives.* Evidence of the first named use of educational tests, to measure the less formal objectives of education, is recorded in the *Tenth Yearbook* of the Department of Superintendence of the National Education Association (400) in which a careful evaluation of available tests for character education is made. The June, 1932, issue of the *Review of Educational Research* (461), likewise, presented important research investigations in the field of tests of personality and character. Tyler (446, 447, 448) clarified the problem of the construction of achievement tests which will really measure the entire range of objectives set up in the educational program. Spencer (432) reported improvement of teaching by means of home-made, non-standardized, diagnostic tests and remedial instruction. Ruch (424), Mann (387), and Smith (431) contributed studies which reveal the relationship between testing and the objectives of education.

3. Evaluating Methods of Teaching

The improvement of educational tests facilitated the careful comparison of various methods of teaching and learning. Studies of this character are so numerous that only a sampling need be given to indicate how tests are used *as basic data in equating groups* and *as measurements of achievement* brought about by the method under study. Zirbes' (467) comparative studies of current practice in reading showed uses of tests as a means of appraising procedures for improving teaching. Broening (324) used tests in literary appreciation as a measure of growth due to a method of teaching literature, as well as reading, intelligence, and appreciation tests for equating the groups of pupils used in the experiment. Coryell (339: 4-5, 25-32, 43) used tests in equating groups and as measures of growth due to the use of the experimental factor in her comparison of intensive and extensive teaching of literature. Some implications of importance are drawn from her experiment regarding the use of objective tests of literature. Raguse (417) analyzed quantitative and qualitative achievement in first-grade reading. Hanna (366) used educational tests in evaluating three methods of problem solving. Newlun (404) utilized objective tests in measuring the values of ability to summarize and achievement in the social studies. Field (351) presented a comparison of reading test scores to prove that extensive individual reading and class reading are desirable procedures to use in teaching of reading in grades three and four.

4. Improvement of Learning

The last ten years has seen an important development in the use of *tests to discover learning difficulty, sources of motivation,* and *uses of self-teaching tests.* This movement has affected the grade placement of items of subjectmatter, the definiteness of objectives, and the use by pupils of self-appraisal and practice tests.

Tests to discover learning difficulty—A current type of study directed toward improvement of learning is that in which tests are formulated for the purpose of investigating the question at issue. Typical examples of such studies are to be found in the field of problem solving in arithmetic. With contrasting sets of material, Washburne and Morphett (460) showed that the children studied succeeded somewhat better when the problems involved familiar situations. In an extensive study of 350,000 problem solutions, Hydle and Clapp (374) examined the effect of eight elements of problem difficulty. In like manner, Monroe (393), with materials selected to suit the purpose, made careful analysis of the nature of pupils' mental processes in solving problems. From another angle Bowman (321) studied the relation between children's success with materials for which they expressed a preference, the *r* between success and reported preference being .56. Wheat (462) measured, with especially selected materials, the differ-

ences in response to problems of the conventional and imaginative type. Kramer (380) investigated the effect of four factors, interest, problem form, vocabulary, and language details, upon sixth-grade children's success in the solution of the verbal arithmetic problem. Among the studies which throw light on the learning difficulty of subjectmatter items are Courtis Standard Practice Tests in Arithmetic (340); Osburn's work in arithmetic (408); the Compass Diagnostic Tests in Arithmetic by Ruch, Knight, Greene, and Studebaker (422); Diagnostic Tests in Arithmetic by Brueckner and others (326). Washburne (457) described a group of experiments to determine the grade placement of several topics in arithmetic; Pressey's test (415) yielded a statistical study of children's errors in sentence structure; Renwick (419) used tests in determining children's difficulties in the study of mensuration; Morton (397, 398, 399) made an analysis of errors in the solution of arithmetic problems; Streitz (435), by using tests, discovered difficulties in arithmetic and their correctives; Brigham (322) made a study of error in the college entrance examination; and Goodenough (360), through the use of tests, worked out a problem of efficiency in learning and the accomplishment ratio.

Sources of motivation have been studied in a more convincing manner through the use of tests. O'Shea's study (409) of the effect of interest of a passage on learning vocabulary used intelligence, reading, comprehension, and vocabulary, and also special objective vocabulary tests. She (409: 44, 49) stated that the significant fact about a pupil's performance is the changes in his score from the original test to the retest, and that the vocabulary tests are as comparable with the standard tests as the standard tests are with each other. Uhl (450) showed the use of standardized materials in arithmetic for diagnosing pupils' methods of work. Symonds (436) used the Charters Diagnostic Language Tests in grade six in six New York public schools and found that test motivation caused learning over and above that which could be explained by practice. The value of the test motivation can be estimated as the equivalent of five sheer repetitions. Book and Norvell (320) discussed the will to learn—an experimental study of incentives which included the use of tests. Curtis and Woods (343) showed the use of new-type tests as a teaching device in science and offered a method of correction which requires the least of the teacher's time and energy.

The Winnetka technic (403) of individualized instruction put renewed emphasis on the use of *tests as a teaching instrument.* Other experimentally determined studies of the effect of practice exercises include such studies as that by Leonard (382), who used objective tests to measure eighth- and ninth-grade pupils' abilities in punctuation and capitalization and practice exercises as a method of improving pupils' ability to write compositions free from errors. His results of the use of practice exercises are statistically convincing both in terms of objective proof-reading tests and in composition writing; the pupil in the experimental group using the practice tests did

almost twice as well as the pupils taught by other methods. Ziegler (466) has convincing objective evidence that test-determined teaching of literature in secondary schools produces greater appreciation of literature as well as improvement in silent reading through the use of practice exercises.

CRITICAL ISSUES

Consideration of present uses of educational tests suggests certain critical issues, some of which are treated elsewhere in this volume, but which are so vital that they justify restressing. Foremost among these is the original "ubiquitous probable error," present from the first, but too frequently forgotten in the enthusiasm of novices. No one has set forth the significance of this item better than Kelley (378). He rightly pointed out the sharp statistical distinction in the reliability between group measurements and individual measurements. There is a tendency for the average user of tests to forget that whereas medians obtained from the tests utilized may have a satisfactory degree of reliability, individual scores in many cases do not. The present tendency in the testing movement in public schools is, at least, distinctly toward greater and greater analysis of learning difficulties as disclosed by test scores. Moreover, the tendency is not alone toward emphasizing individual pupils' *total* scores but toward utilizing individual pupils' scores on items in a test. Now it is well known that a pupil's reaction to a single question is highly unreliable. Yet this fact is often overlooked. Thus certainty that a pupil actually does not know that $7 + 8 = 15$ cannot be safely inferred from a single trial, but this is probably frequently being done. The need here is for longer and more analytical tests. Some progress has been made along this line. Examples are the Compass Arithmetic Tests (422) and Osburn's work in arithmetic (408).

Another critical issue that has been vigorously raised by Kelley is the fallacious assumption that intelligence and achievement tests measure essentially different attributes. If his contention that the typical intelligence test measures 90 percent of the identical traits measured by a good achievement test is true, it vitiates many of the rather common procedures of comparing a pupil's achievement score with his intelligence score. In particular, as he points out, because of the almost total absence of zero points in all tests, quotient technics are particularly treacherous.

Another critical issue which should challenge persons working in the field of educational measurement deals not so much with methods as with objectives. Broadly speaking the tools of educational measurement have been utilized chiefly in the attempt to improve educational methods. It seems to this reviewer, at least, that a far more urgent need at the present juncture is for research in the field of objectives or aims in the field of education. Tyler (446) did pioneer work in this direction in attempting to enlarge the scope of tests in certain fields to measure more adequately the *total objectives* of courses offered than has been done in the past. Unquestionably the

most urgent problems in education at present are sociological in nature, and while the application of measurement technics to sociological problems may be vastly more difficult than are the problems of methods of teaching, the issue cannot be escaped on this ground. The technics of educational measurement should aim to contribute as much to the problem of what type of training the present-day school graduate should receive as they have in the appraisal of an improvement in the quality of that training. This may involve totally new technics and years of effort, but the challenge cannot be avoided.

In the use of tests in typical school systems, an important principle apparently is that of programming tests in some sort of systematically recurring schedules as opposed to sporadic testings. Enormously greater gains can thus be obtained from test results due to what might be called unearned increments. Systematically recurring testing programs make possible greatly enriched programs of supervision and instructional research that are not possible on the basis of sporadic tests. The work of the Baltimore Bureau of Research (433) is an example of this type of procedure. In this school system educational tests are given to every pupil in the system at the beginning of each term in accordance with long range planning.

Another point, though not strictly a matter of testing but rather a matter of principles of education, may be mentioned. It is the levelling process which seems too frequently to follow testing programs. Kelley (377) admirably set forth arguments against levelling of pupils' abilities after discovering individual idiosyncrasies and arguments for the preservation of these idiosyncrasies. This reviewer is in agreement with his point of view. Individual differences, while so troublesome in neat mechanical schemes of school administration, are none the less unquestionably our greatest assets and for this reason should be preserved in-so-far as is practical.

Another critical issue is the need for mechanization of test procedures to reduce the time and labor cost. To make full use of even the tests already available requires more time and energy than is generally available under present-day school conditions. Since a large part of the work, particularly that of scoring and gross tabulating, is essentially mechanical in nature, there is great need for mechanical means for carrying out this work so as to relieve teachers and test technicians for more interpretative work. It is an encouraging sign to note a number of movements in this direction. As the demand becomes stronger, unquestionably machinery will be developed for doing much of this work. Several examples of efforts along this line may be noted. Thus Clapp and Young's (336) contribution of the carbon offset scoring process, and Toops's (444) envelope scheme for writing all answers of a test on a single sheet to economize test booklets tend to reduce labor. In the computing field, a number of technics have been reported (346, 348, 413) whereby correlations and other computations may be computed by means of standard computing machinery. Hull (373) described

the most elaborate computing machine for making the almost interminable calculations utilized in the partial and multiple correlation technics for combining tests according to optimum weights. A considerable number of charts, tables, and slide rules have recently appeared. Among these may be mentioned Dunlap and Kurtz's *Handbook of Statistical Nomographs* (346), the Otis Correlation Chart (410), the Universal Percentile Graph (412), the Kelley Correlation Chart (377), the Stenquist Teachers Class Analysis Charts (434).

The heaviest load of the testing program still remains, as before, the basic scoring of tests. Unofficial reports indicate that a number of persons are at work in the development of machines for doing this work, and it may confidently be hoped that in the near future much of this burden may be accomplished by this means. Until such facilities are provided, the full value of tests can never be obtained on any large scale. The diagnostic values possible by taking account of individual performances on individual items under the proper conditions are at present, for the most part, lost because of the huge labor load involved.

BIBLIOGRAPHY ON EDUCATIONAL TESTS AND THEIR USES

Chapter I. Basic Considerations

1. Carnegie Foundation for the Advancement of Teaching. *Study of the Relations of Secondary and Higher Education in Pennsylvania.* Progress Reports III and IV. New York: the Foundation, 1931.
2. Hawkes, H. E. "Cooperative Test Service." *Educational Record* 12: 30-38; January, 1931.
3. Hawkes, H. E., chairman, and others. *Personnel Methods.* Educational Record Supplement No. 8. Washington, D. C.: American Council on Education, July, 1928. 68 p.
4. Johnston, J. B., chairman. "The 1932 College Sophomore Testing Program; a Report by the Advisory Committee on College Testing." *Educational Record* 13: 290-343; October, 1932.
5. Lindquist, E. F. "The Form of the American History Examination of the Cooperative Test Service." *Educational Record* 12: 459-75; October, 1931.
6. McConn, Max. "The Carnegie Foundation's Study of Secondary and Higher Education in Pennsylvania." *Bulletin of the American Association of Collegiate Registrars* 5: 43-54; January, 1930.
7. McConn, Max. "The Co-operative Test Service." *Journal of Higher Education* 2: 225-32; May, 1931.
8. Wood, Ben D. "The Cooperative Test Service." *Educational Record* 12: 244-52; July, 1931.
9. Wood, Ben D. "The Structure and Content of the Comprehensive Examination for College Sophomores." *Recent Trends in American College Education.* Proceedings of the Institute for Administrative Officers of Higher Institutions, Vol. 3. Chicago: University of Chicago Press, 1931. Chapter 20, p. 190-207.

Chapter II. Recent Developments in the Selection of Test Items

10. Ashbaugh, Ernest J. *The Iowa Spelling Scales; Their Derivation, Uses, and Limitations.* Journal of Educational Research Monographs, No. 3. Bloomington, Ill.: Public School Publishing Co., 1922. 144 p. (Doctor's thesis).
11. Ayres, Leonard P. *A Measuring Scale for Ability in Spelling.* New York: Russell Sage Foundation, 1915. 58 p.
12. Ballou, Frank W. *Scales for the Measurement of English Compositions.* Harvard Bulletins in Education, No. 2. Cambridge, Mass.: Harvard University Press, 1914. 93 p.
13. Barr, A. S. "Measurements and Progressive Education (Editorial)." *Journal of Educational Research* 22: 317-19; November, 1930.
14. Bayles, Ernest E., and Bedell, Ralph C. "A Study of Comparative Validity as Shown by a Group of Objective Tests." *Journal of Educational Research* 23: 8-16; January, 1931.
15. Brinkley, Sterling G. *Values of New Type Examinations in the High School.* Contributions to Education, No. 161. New York: Teachers College, Columbia University, 1924. 121 p.
16. Buckingham, B. R. *Spelling Ability: Its Measurement and Distribution.* Contributions to Education, No. 59. New York: Teachers College, Columbia University, 1913. 116 p.
17. Clark, E. L. "A Method of Evaluating the Units of a Test." *Journal of Educational Psychology* 19: 263-65; April, 1928.
18. Cook, Walter W. *The Measurement of General Spelling Ability Involving Controlled Comparison between Techniques.* University of Iowa Studies in Education, Vol. 6, No. 6. Iowa City: the University, 1932. 112 p.

19. COREY, STEPHEN M. "The Effect of Weighting Exercises in a New Type Examination." *Journal of Educational Psychology* 21: 383-85; May, 1930.
20. COURTIS, S. A. *The Measurement of Classroom Products.* New York: General Education Board, 1920.
21. COURTIS, S. A. *Standard Research Test in Arithmetic.* Detroit, Mich.: the Author.
22. COURTIS, S. A. *Standard Research Tests in Reading.* Detroit, Mich.: the Author.
23. DOLCH, EDWARD WILLIAM. "Sampling of Reading Matter." *Journal of Educational Research* 22: 213-15; October, 1930.
24. DOUGLASS, HARL R., and SPENCER, PETER L. "Is It Necessary to Weight Measures in Standard Tests?" *Journal of Educational Psychology* 14: 109-12; February, 1923.
25. EURICH, ALVIN C. "Four Types of Examinations Compared and Evaluated." *Journal of Educational Psychology* 22: 268-78; April, 1931.
26. GATES, ARTHUR I. "The Correlations of Achievement in School Subjects with Intelligence Tests and Other Variables." *Journal of Educational Psychology* 13: 129-39; March, 1922.
27. GATES, ARTHUR I. "A Test of Ability in the Pronunciation of Words." *Teachers College Record* 26: 205-19; November, 1924.
28. GATES, ARTHUR I. "The True-False Test as a Measure of Achievement in College Courses." *Journal of Educational Psychology* 12: 276-87; May, 1921.
29. KELLEY, TRUMAN LEE. "A Communication Concerning the Difficulty of Achievement Test Items." *Journal of Educational Research* 22: 309-14; November, 1930.
30. KELLEY, TRUMAN LEE. "Note on the Reliability of a Test: A Reply to Dr. Crum's Criticism." *Journal of Educational Psychology* 15: 193-204; April, 1924.
31. KELLEY, TRUMAN LEE; RUCH, G. M.; and TERMAN, LEWIS M. *The Stanford Achievement Test.* Yonkers-on-Hudson, N. Y.: World Book Co., 1922. "Manual of Directions" 8-14.
32. LENTZ, THEODORE F.; HIRSHSTEIN, BERTHA; and FINCH, F. H. "Evaluation of Methods of Evaluating Tests Items." *Journal of Educational Psychology* 23: 344-50; May, 1932.
33. LEWIS, E. E. *Scales for Measuring Special Types of English Composition.* Yonkers-on-Hudson, N. Y.: World Book Co., 1921. 144 p.
34. LINDQUIST, E. F., and ANDERSON, H. R. "Objective Testing in World History." *Historical Outlook* 21: 115-22; March, 1930.
35. LYON, ELVA ANNE. "Objective Measurements in English." *Educational Research Bulletin (Ohio State University)* 9: 481-89; November 19, 1930.
36. MCCALL, WILLIAM A., and OTHERS. "Construction of Multi-Mental Scale." *Teachers College Record* 27: 394-415; January, 1926.
37. MCCALL, WILLIAM A. *How to Measure in Education.* New York: Macmillan Co., 1922. Chapter 7, "Preparation and Validation of Test Material," p. 195-226.
38. MONROE, WALTER S. *Directing Learning in the High School.* Garden City, N. Y.: Doubleday, Page and Co., 1927. Chapters 15-16, p. 473-538.
39. MONROE, WALTER S. "An Experimental and Analytical Study of Woody's Arithmetic Scales, Series B." *School and Society* 6: 412-20; October 6, 1917.
40. MONROE, WALTER S., and SOUDERS, LLOYD B. *The Present Status of Written Examinations and Suggestions for Their Improvement.* University of Illinois Bulletin, Vol. 21, No. 13, Educational Research Bulletin, No. 17. Urbana, Ill.: the University, 1923. 77 p.
41. ODELL, CHARLES W. "Further Data Concerning the Effect of Weighting Exercises in the New-Type Examinations." *Journal of Educational Psychology* 22: 700-4; December, 1931.
42. ODELL, CHARLES W. *Scales for Rating Pupils' Answers to Nine Types of Thought Questions in General Science.* Urbana, Ill.: Bureau of Educational Research, University of Illinois, 1927. 20 p.
43. OSBURN, W. J. "The Improvement of the Essay Examination." To be published in *Journal of Educational Research.*
44. OTIS, ARTHUR S. "The Reliability of Spelling Scales, Involving a 'Deviation Formula' for Correlation." *School and Society* 4: 750-56, 793-96; November 11-18, 1916.
45. PATERSON, DONALD G. *The Preparation and Use of New-Type Examinations.* Yonkers-on-Hudson, N. Y.: World Book Co., 1926. 87 p.

46. PETERS, CHARLES C., and ALTMAN, JOHN E. "A Critical Study of the Content of Standardized Tests in American History." *Journal of Educational Research* 23: 153-61; February, 1931.
47. PETERS, CHARLES C. "The Relation of Standardized Tests to Educational Objectives." *Objectives of Education.* Second Yearbook. National Society for the Study of Educational Sociology. New York: Teachers College, Columbia University, 1929. p. 148-59.
48. POTTHOFF, EDWARD F., and BARNETT, NEVILLE E. "A Comparison of Marks Based upon Weighted and Unweighted Items in a New-Type Examination." *Journal of Educational Psychology* 23: 92-98; February, 1932.
49. PRESSEY, SIDNEY L., and PRESSEY, LUELLA C. *Introduction to the Use of Standard Tests.* Yonkers-on-Hudson, N. Y.: World Book Co., 1922. p. 8-11.
50. RUCH, G. M. *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Co., 1929. p. 31-39.
51. RUCH, G. M., and STODDARD, GEORGE D. *Tests and Measurements in High School Instruction.* Yonkers-on-Hudson, N. Y.: World Book Co., 1927. p. 48-51, 301-28.
52. RUSSELL, CHARLES. *Classroom Tests.* Boston: Ginn and Co., 1926. p. 14-16.
53. SEASHORE, CARL EMIL. *The Psychology of Musical Talent.* New York: Silver, Burdett and Co., 1919. p. 7-8.
54. SIMS, VERNER MARTIN. "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating." *Journal of Educational Research* 24: 216-23; October, 1931.
55. SONES, W. W. D., and HARRY, DAVID P. *Sones-Harry High-School Achievement Test.* Yonkers-on-Hudson, N. Y.: World Book Co., 1929. Forms A and B, each 24 p.
56. STONE, C. W. *Arithmetical Abilities and Some Factors Determining Them.* Contributions to Education, No. 19. New York: Teachers College, Columbia University, 1908. 101 p.
57. SYMONDS, PERCIVAL M. "Choice of Items for a Test on the Basis of Difficulty." *Journal of Educational Psychology* 20: 481-93; October, 1929.
58. SYMONDS, PERCIVAL M. "Factors Influencing Test Reliability." *Journal of Educational Psychology* 19: 73-87; February, 1928.
59. TALBOTT, E. O., and RUCH, G. M. "Minor Studies on Objective Examination Methods. II. The Theory of Sampling as Applied to Examinations." *Journal of Educational Research* 20: 199-206; October, 1929.
60. THORNDIKE, EDWARD L. "Means of Measuring School Achievement in Spelling." *Educational Administration and Supervision* 1: 306-12; May, 1915.
61. THORNDIKE, EDWARD L. "The Measurement of Ability in Reading." *Teachers College Record* 15: 1-71; September, 1914.
62. THORNDIKE, EDWARD L., and OTHERS. *The Measurement of Intelligence.* New York: Teachers College, Columbia University, 1927. Chapters 2-6.
63. THORNDIKE, EDWARD L. *Preliminary Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People.* New York: Teachers College, Columbia University.
64. THURSTONE, THELMA G. "The Difficulty of a Test and Its Diagnostic Value." *Journal of Educational Psychology* 23: 335-43; May, 1932.
65. TOOPS, HERBERT A.; ADKINS, DOROTHY; and MEYERS, L. *A Follow-up Investigation of a Study Performance Test with a View to Supplementing the Intelligence Test.* Ohio College Association Bulletin, No. 66. Columbus: Ohio State University.
66. TOOPS, HERBERT A., and ROYER, E. B. *Predicting Soldier's School Marks: A Problem in the Selection of Tests.* Ohio College Association Bulletin, No. 80. Columbus: Ohio State University.
67. TRABUE, MARION R. *The Nassau County Supplement to the Hillegas Scale.* New York: Teachers College, Columbia University.
68. TYLER, RALPH W. "A Generalized Technique for Constructing Achievement Tests." *Educational Research Bulletin (Ohio State University)* 10: 199-208; April 15, 1931.
69. TYLER, RALPH W. "Making a Cooperative Test Service Effective." *Educational Research Bulletin (Ohio State University)* 11: 287-92; May 25, 1932.

70. TYLER, RALPH W. "Measuring Ability to Infer." *Educational Research Bulletin (Ohio State University)* 9: 475-80; November 19, 1930.
71. TYLER, RALPH W. "Measuring the Results of College Instruction." *Educational Research Bulletin (Ohio State University)* 11: 253-60; May 11, 1932.
72. TYLER, RALPH W. "A Test of Skill in Using a Microscope." *Educational Research Bulletin (Ohio State University)* 9: 493-96; November 19, 1930.
73. VINCENT, LEONA. *A Study of Intelligence Test Elements.* Contributions to Education, No. 152. New York: Teachers College, Columbia University, 1924. 36 p.
74. WHELDEN, CHESTER H., and DAVIES, F. J. J. "A Method for Judging the Discrimination of Individual Questions on True-False Examinations." *Journal of Educational Psychology* 22: 290-306; April, 1931.
75. WILSON, GUY M. "The Purpose of a Standardized Test in Spelling." *Journal of Educational Research* 20: 319-26; December, 1929.
76. WILSON, WILLIAM R.; WELSH, G.; and GULLIKSEN, HAROLD. "An Evaluation of Some Information Questions." *Journal of Applied Psychology* 8: 206-14; June, 1924.
77. WOOD, BEN D. *Measurement in Higher Education.* Yonkers-on-Hudson, N. Y.: World Book Co., 1923. Chapters 7-8, p. 141-214.
78. WOODY, CLIFFORD. *Measurements of Some Achievements in Arithmetic.* Contributions to Education, No. 80. New York: Teachers College, Columbia University, 1916. 63 p.

Chapter III. Recent Developments in Statistical Procedures

79. ADAMS, EUNICE. *The Comparative Reliability of Eight Arithmetic Tests.* Unpublished master's thesis, University of California, 1929. 37 p.
80. ANDERSON, L. DEWEY, and TOOPS, HERBERT A. "A New Apparatus for Plotting and a Checking Method for Solving Large Numbers of Intercorrelations." *Journal of Educational Psychology* 19: 650-57, December, 1928; 20: 36-43, January, 1929.
81. BATHURST, J. E. "A Partial Correlation Scheme." *Journal of Applied Psychology* 11: 155-64; April, 1927.
82. BLISS, ELISHA F. "The Difficulty of an Item." *Journal of Educational Psychology* 20: 63-66; January, 1929.
83. BLISS, ELISHA F. "Theories Underlying the Statistical Determination of Credit to Be Allotted Item Responses." *Journal of Educational Psychology* 19: 584-85; November, 1928.
84. BRIGHAM, CARL C. *Study of American Intelligence.* Princeton, N. J.: Princeton University Press, 1923. 210 p.
85. BRINKLEY, STERLING G. *Values of New Type Examinations in the High School.* Contributions to Education, No. 161. New York: Teachers College, Columbia University, 1924. 121 p.
86. BRINKMEIER, INA HILL, and KEYS, N. "Circumstantiality as a Factor in Guessing on True-False Examinations." *Journal of Educational Psychology* 21: 681-94; December, 1930.
87. BRINKMEIER, INA HILL, and RUCH, G. M. "Minor Studies on Objective Examination Methods. III. Specific Determiners in True-False Statements." *Journal of Educational Research* 22: 110-18; September, 1930.
88. BRINKMEIER, INA HILL. "Minor Studies on Objective Examination Methods. IV. Sentence Length as a Specific Determiner in True-False Statements." *Journal of Educational Research* 22: 203-5; October, 1930.
89. BROOM, M. EUSTACE; DOUGLAS, JOSEPHINE; and RUDD, MARION. "On the Validity of Silent Reading Tests." *Journal of Applied Psychology* 15: 35-38; February, 1931.
90. BROWN, WILLIAM. "Some Experimental Results in the Correlation of Mental Abilities." *British Journal of Psychology* 3: 296-322; October, 1910.

91. Chapman, J. Crosby, and Chapman, Daisy Rogers. *Trade Tests.* New York: Henry Holt and Co., 1921. 435 p.
92. Chapman, J. Crosby. "The Unreliability of the Difference between Intelligence and Educational Ratings." *Journal of Educational Psychology* 14: 103-8; February, 1923.
93. Clark, E. L. "A Method of Evaluating the Units of a Test." *Journal of Educational Psychology* 19: 263-65; April, 1928.
94. Cleeton, Glen U. "Optimum Difficulty of Group Test Items." *Journal of Applied Psychology* 10: 327-40; September, 1926.
95. Corey, Stephen M. "The Effect of Weighting Exercises in a New Type of Examination." *Journal of Educational Psychology* 21: 383-85; May, 1930.
96. Courtis, S. A. "Maturation Units for the Measurement of Growth." *School and Society* 30: 683-90; November 16, 1929.
97. Crum, W. L. "Note on the Reliability of a Test, with Special Reference to the Examinations Set by the College Entrance Board." *American Mathematical Monthly* 30: 296-301; September-October, 1923.
98. Cureton, Edward E., and Dunlap, Jack W. "A Nomograph for Estimating a Reliability Coefficient by the Spearman-Brown Formula and for Computing Its Probable Error." *Journal of Educational Psychology* 21: 68-69; January, 1930.
99. Cureton, Edward E., and Dunlap, Jack W. "Nomograph for Estimating the Reliability of a Test in One Range of Talent When Its Reliability Is Known in Another Range." *Journal of Educational Psychology* 20: 537-38; October, 1929.
100. Cureton, Edward E. "Note on the Computation of the Rank-Difference Correlation Coefficient." *Journal of Educational Psychology* 18: 627-30; December, 1927.
101. Current, W. F., and Ruch, G. M. "Further Studies on the Reliability of Reading Tests." *Journal of Educational Psychology* 17: 476-81; October, 1926.
102. Dodd, Stuart C. "A Correlation Machine." *Industrial Psychology* 1: 46-58; January, 1926.
103. Douglass, H. R., and Huffaker, C. L. "Correlation between Intelligence Quotient and Accomplishment Quotient." *Journal of Applied Psychology* 13: 76-80; February, 1929.
104. Douglass, H. R., and Spencer, P. L. "Is It Necessary to Weight Exercises in Standard Tests?" *Journal of Educational Psychology* 14: 109-12; February, 1923.
105. Douglass, H. R., and Cozens, F. W. "On Formula for Estimating the Reliability of Test Batteries." *Journal of Educational Psychology* 20: 369-77; May, 1929.
106. Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas.* Yonkers-on-Hudson, N. Y.: World Book Co., 1932. 163 p.
107. Dvorak, August, and Jensen, Myra. *Linear and Non-Linear Correlation Chart.* New York: Longmans, Green, and Co., 1931.
108. Dvorak, August. "A Simplified Computation of Non-Linear Correlation." *Journal of Educational Research* 25: 99-104; February, 1932.
109. Edgerton, Harold A. "An Abac for Finding the Standard Error of a Proportion and the Standard Error of the Difference of Proportions." *Journal of Educational Psychology* 18: 127-28; 350; February and May, 1927.
110. Edgerton, Harold A., and Toops, Herbert A. "A Formula for Finding the Average Intercorrelation Coefficient of Unranked Raw Scores without Solving any of the Individual Intercorrelations." *Journal of Educational Psychology* 19: 131-38; February, 1928.
111. Edgerton, Harold A. "A Graphic Method of Finding Standard Errors and Probable Errors of Differences." *Journal of Educational Psychology* 23: 56-57; January, 1932.
112. Edgerton, Harold A. "A Table for Finding the Probable Error of R Obtained by Use of the Spearman-Brown Formula (n = 2)." *Journal of Applied Psychology* 14: 296-302; June, 1930.
113. Edgerton, Harold A., and Toops, Herbert A. "A Table for Predicting the Validity and Reliability Coefficients of a Test When Lengthened." *Journal of Educational Research* 18: 225-34; October, 1928.
114. Edgerton, Harold A., and Paterson, D. G. "Table of Standard Errors and Probable Errors of Percentages for Varying Numbers of Cases." *Journal of Applied Psychology* 10: 378-91; September, 1926.

115. EURICH, ALVIN C. "Four Types of Examinations Compared and Evaluated." *Journal of Educational Psychology* 22: 268-78; April, 1931.
116. FARNSWORTH, PAUL R. "Concerning So-called Group Effects." *Pedagogical Seminary and Journal of Genetic Psychology* 35: 587-94; December, 1928.
117. FARNSWORTH, PAUL R. "The Spearman-Brown Prophecy Formula and the Seashore Tests." *Journal of Educational Psychology* 19: 586-88; November, 1928.
118. FORAN, T. G. *The Meaning and Limitations of Scores, Norms, and Standards in Educational Measurement.* Catholic University of America, Educational Research Bulletins, Vol. 3, No. 2. Washington, D. C.: Catholic Education Press, 1928. p. 16-19.
119. FORAN, T. G. "A Note on Methods of Measuring Reliability." *Journal of Educational Psychology* 22: 383-87; May, 1931.
120. FOSTER, R. R., and RUCH, G. M. "On Corrections for Chance in Multiple-Response Tests." *Journal of Educational Psychology* 18: 48-51; January, 1927.
121. FRANZEN, RAYMOND H. "The Accomplishment Quotient." *Teachers College Record* 21: 432-40; November, 1920.
122. FREEMAN, FRANK N. *Mental Tests.* Boston: Houghton Mifflin Co., 1926. p. 247-56.
123. FREEMAN, FRANK S. "Power and Speed: Their Influence upon Intelligence Test Scores." *Journal of Applied Psychology* 12: 631-35; December, 1928.
124. GALTON, FRANCIS. "Grades and Deviates." *Biometrika* 5: 400-6; June, 1907.
125. GARRETT, HENRY E. "A Modification of Tolley and Ezekiel's Method of Handling Multiple Correlation Problems." *Journal of Educational Psychology* 19: 45-49; January, 1928.
126. GATES, ARTHUR I. "An Experimental and Statistical Study of Reading and Reading Tests." *Journal of Educational Psychology* 12: 303-14; 378-91; 445-64; September-November, 1921.
127. GORDON, KATE. "Group Judgments in the Field of Lifted Weights." *Journal of Experimental Psychology* 7: 398-400; October, 1924.
128. GRIFFIN, HAROLD D. "Nomogram for Checking the Reliability of Test Scores." *Journal of Applied Psychology* 14: 609-11; December, 1930.
129. GRIFFIN, HAROLD D. "Nomograms for Correcting Simple and Multiple Correlation Coefficients." *Journal of the American Statistical Association* 25: 316-19; September, 1930.
130. HERRING, JOHN P. "The Reliability of Accomplishment Differences." *Journal of Educational Psychology* 15: 530-38; November, 1924.
131. HOLZINGER, KARL J. "An Analysis of the Errors in Mental Measurement." *Journal of Educational Psychology* 14: 278-88; May, 1923.
132. HOLZINGER, KARL J. *Correlation Chart.* Chicago: the Author (University of Chicago).
133. HOLZINGER, KARL J., and CLAYTON, BLYTHE. "Further Experiments in the Application of Spearman's Prophecy Formula." *Journal of Educational Psychology* 16: 289-99; May, 1925.
134. HOLZINGER, KARL J. "Note on the Use of Spearman's Prophecy Formula for Reliability." *Journal of Educational Psychology* 14: 302-5; May, 1923.
135. HOLZINGER, KARL J. "On Scoring Multiple Response Tests." *Journal of Educational Psychology* 15: 445-47; October, 1924.
136. HOLZINGER, KARL J. "Some Comments on Professor Thurstone's Method of Determining the Scale Values of Test Items." *Journal of Educational Psychology* 19: 112-26; February, 1928.
137. HOLZINGER, KARL J. *Statistical Tables for Students in Education and Psychology.* 3d ed. Chicago: University of Chicago Press. 1931. 101 p.
138. HOLZINGER, KARL J. *Tables of the Probable Error of the Coefficient of Correlation as Found by the Product Moment Method.* London: Cambridge University Press, 1925.
139. HUFFAKER, CARL L. "A Contribution to the Technique of Partial Correlation." *Journal of Applied Psychology* 7: 135-42; June, 1923.
140. HUFFAKER, CARL L. "The Probable Error of the Accomplishment Quotient." *Journal of Educational Psychology* 21: 550-51; October, 1930.
141. HULL, CLARK L. *Aptitude Testing.* Yonkers-on-Hudson, N. Y.: World Book Co., 1928. 536 p.

142. HULL, CLARK L. "An Automatic Machine for Making Multiple Aptitude Forecasts." *Journal of Educational Psychology* 16: 593-98; December, 1925.
143. KELLEY, TRUMAN LEE. "The Applicability of the Spearman-Brown Formula for the Measurement of Reliability." *Journal of Educational Psychology* 16: 300-3; May, 1925.
144. KELLEY, TRUMAN LEE. *Chart to Faciliatate the Calculation of Partial Coefficients of Correlation and Regression Equations.* Stanford University, School of Education, Special Monograph No. 1. Stanford University, Calif.: the University, 1921. 24 p.
145. KELLEY, TRUMAN LEE. *Correlation Chart.* Stanford University, Calif.: Stanford University Book Store.
146. KELLEY, TRUMAN LEE, and MCNEMAR, QUINN. "Doolittle versus the Kelley-Salisbury Iteration Method for Computing Multiple Regression Coefficients." *Journal of the American Statistical Association* 24: 164-69; June, 1929.
147. KELLEY, TRUMAN LEE. *Educational Guidance.* Contributions to Education, No. 71. New York: Teachers College, Columbia University, 1914. 116 p.
148. KELLEY, TRUMAN LEE. *Interpretation of Educational Measurements.* Yonkers-on-Hudson, N. Y.: World Book Co., 1927. 363 p.
149. KELLEY, TRUMAN LEE, and SALISBURY, FRANK S. "An Iteration Method for Determining Multiple Correlation Constants." *Journal of the American Statistical Association* 21: 282-92; September, 1926.
150. KELLEY, TRUMAN LEE. "A New Method for Determining the Significance of Differences in Intelligence and Achievement Scores." *Journal of Educational Psychology* 14: 321-33; September, 1923.
151. KELLEY, TRUMAN LEE. "The Reliability of Test Scores." *Journal of Educational Research* 3: 370-79; May, 1921.
152. KELLEY, TRUMAN LEE. *Statistical Method.* New York: Macmillan Co., 1923. 390 p.
153. KELLEY, TRUMAN LEE. *Tables: to Facilitate the Calculation of Partial Coefficients of Correlation and Regression Equations.* Bulletin of the University of Texas, No. 27. Austin, Tex.: the University, 1916. 53 p.
154. LANIER, LYLE H. "Prediction of the Reliability of Mental Tests and Tests of Special Abilities." *Journal of Experimental Psychology* 10: 69-113; April, 1927.
155. LENTZ, THEODORE F.; HIRSHSTEIN, BERTHA; and FINCH, F. H. "Evaluation of Methods of Evaluating Test Items." *Journal of Educational Psychology* 23: 344-50; May, 1932.
156. LENTZ, THEODORE F. "Utilizing Opinion for Character Measurement." *Journal of Social Psychology* 1: 536-42; November, 1930.
157. LINCOLN, E. A. "The Unreliability of Reliability Coefficients." *Journal of Educational Psychology* 23: 11-14; January, 1932.
158. LINDQUIST, E. F. "Factors Determining Reliability of Test Norms." *Journal of Educational Psychology* 21: 512-20; October, 1930.
159. LONGSTAFF, H. P., and PORTER, J. P. "Speed and Accuracy as Factors in Objective Tests in General Psychology." *Journal of Applied Psychology* 12: 636-42; December, 1928.
160. MANGOLD, SISTER MARY CECILIA. *Methods for Measuring the Reliability of Tests.* Catholic University of America, Educational Research Bulletins, Vol. 2, No. 8. Washington, D. C.: Catholic Education Press, 1927. 32 p.
161. MASTERS, H. V., and UPSHALL, C. C. "Tables of Probable Errors for Certain Interpercentile Ranges." *Journal of Educational Psychology* 23: 287-90; April, 1932.
162. MATHEWS, CHESTER O. "The Effect of Position of Printed Response Words upon Children's Answers to Questions in Two-Response Types of Tests." *Journal of Educational Psychology* 18: 445-57; October, 1927.
163. MATHEWS, CHESTER O. "The Effect of the Order of Printed Response Words on an Interest Questionnaire." *Journal of Educational Psychology* 20: 128-34; February, 1929.
164. MCCALL, WILLIAM A., and OTHERS. "Construction of a Multi-Mental Scale." *Teachers College Record* 27: 394-415; January, 1926.
165. MCCALL, WILLIAM A. *How to Measure in Education.* New York: Macmillan Co., 1922. 416 p.

166. McCall, William A. "Proposed Uniform Method of Scale Construction." *Teachers College Record* 22: 31-51; January, 1921.
167. McCrory, John R. "The Reliability of the Accomplishment Quotient." *Journal of Educational Research* 25: 27-39; January, 1932.
168. Mendenhall, R. M., and Warren, Richard. "Computing Statistical Coefficients from Punched Cards." *Journal of Educational Psychology* 21: 53-62; January, 1930.
169. Meyer, Henry W. *The Effect of Printed Response Words upon Children's Answers to Questions in Two-Response Types of Tests.* Unpublished master's thesis, University of California, 1930. 38 p.
170. Monroe, Walter S. *A Critical Study of Certain Silent Reading Tests.* University of Illinois Bulletin, Vol. 19, No. 22, Educational Research Bulletin, No. 8. Urbana, Ill.: the University, 1922. 52 p.
171. Monroe, Walter S.; DeVoss, James C.; and Kelly, Frederick J. *Educational Tests and Measurements.* Boston: Houghton Mifflin Co., 1917. 309 p.
172. Monroe, Walter S.; DeVoss, James C.; and Kelly, Frederick J. *Educational Tests and Measurements.* Rev. ed. Boston: Houghton Mifflin Co., 1924. 521 p.
173. Monroe, Walter S., and Buckingham, B. R. *Illinois Examination; Teachers Handbook.* Urbana, Ill.: Bureau of Educational Research, University of Illinois, 1920. 32 p.
174. Monroe, Walter S. "Improvement of Instruction Through the Use of Educational Tests." *Journal of Educational Research* 1: 96-102; February, 1920.
175. Monroe, Walter S. *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Co., 1923. 364 p.
176. Morley, Clyde A. "The Reliability of the Achievement Quotient." *Journal of Educational Psychology* 21: 351-60; May, 1930.
177. Mosher, Raymond M. "A Further Note on the Reliability of Reading Tests." *Journal of Educational Psychology* 19: 272-74; April, 1928.
178. Muenzinger, Karl F. "Critical Note on the Reliability of a Test." *Journal of Educational Psychology* 18: 424-28; September, 1927.
179. Odell, Charles W. *A Critical Study of Measures of Achievement Relative to Capacity.* University of Illinois Bulletin, Vol. 26, No. 29, Educational Research Bulletin, No. 45. Urbana, Ill.: the University, 1929. p. 44-46.
180. Odell, Charles W. "Further Data Concerning the Effect of Weighting Exercises in New-Type Examinations." *Journal of Educational Psychology* 22: 700-4; December, 1931.
181. Orleans, Jacob S. "Correlation without Plotting." *Journal of Educational Psychology* 18: 310-17; May, 1927.
182. Otis, Arthur S. *Correlation Chart.* Yonkers-on-Hudson, N. Y.: World Book Co.
183. Paterson, Donald G., and Langlie, T. A. "Empirical Data on the Scoring of True-False Tests." *Journal of Applied Psychology* 9: 339-48; December, 1925.
184. Peak, Helen, and Boring, Edwin G. "The Factor of Speed in Intelligence." *Journal of Experimental Psychology* 9: 71-94; April, 1926.
185. Peatman, John Gray. "The Influence of Weighted True-False Test Scores on Grades." *Journal of Educational Psychology* 21: 143-47; February, 1930.
186. Peters, Charles C., and Wykes, Elizabeth Crossley. "Simplified Methods for Computing Regression Coefficients and Partial and Multiple Correlations." *Journal of Educational Research* 23: 383-93, May, 1931; 24: 44-52, June, 1931.
187. Pintner, Rudolf, and Marshall, H. R. "A Combined Mental-Educational Survey." *Journal of Educational Psychology* 12: 32-43; January, 1921.
188. Popenoe, Herbert. "A Report of Certain Significant Deficiencies of the Accomplishment Quotient." *Journal of Educational Research* 16: 40-47; June, 1927.
189. Potthoff, Edward F., and Barnett, Neville E. "A Comparison of Marks Based upon Weighted and Unweighted Items in a New-Type Examination." *Journal of Educational Psychology* 23: 92-98; February, 1932.
190. Rand, Gertrude. "A Discussion of the Quotient Method of Specifying Test Results." *Journal of Educational Psychology* 16: 599-618; December, 1925.
191. Remmers, H. H.; Shock, N. W.; and Kelly, E. L. "An Empirical Study of the Validity of the Spearman-Brown Formula as Applied to the Purdue Rating Scale." *Journal of Educational Psychology* 18: 187-95; March, 1927.

192. REMMERS, H. H. "The Equivalence of Judgments to Test Items in the Sense of the Spearman-Brown Formula." *Journal of Educational Psychology* 22: 66-71; January, 1931.
193. REMMERS, H. H., and OTHERS. "An Experimental Study of the Relative Difficulty of True-False, Multiple-Choice, and Incomplete-Sentence Types of Examination Questions." *Journal of Educational Psychology* 14: 367-72; September, 1923.
194. ROSENOW, CURT. *The Analysis of Mental Functions.* Psychological Monographs, Vol. 24, No. 5. Whole No. 106. Princeton, N. J.: Psychological Review Co., 1917. 43 p.
195. RUCH, G. M. "The Achievement Quotient Technique." *Journal of Educational Psychology* 14: 334-43; September, 1923.
196. RUCH, G. M., and MEYER, STANTON H. "Comparative Merits of Physics Tests." *School Science and Mathematics* 31: 676-80; June, 1931.
197. RUCH, G. M., and STODDARD, GEORGE D. "Comparative Reliabilities of Five Types of Objective Examinations." *Journal of Educational Psychology* 16: 89-103; February, 1925.
198. RUCH, G. M., and CHARLES, JOHN W. "A Comparison of Five Types of Objective Tests in Elementary Psychology." *Journal of Applied Psychology* 12: 398-403; August, 1928.
199. RUCH, G. M., and DEGRAFF, MARK H. "Corrections for Chance and 'Guess' vs. 'Do not Guess' Instructions in Multiple-Response Tests." *Journal of Educational Psychology* 17: 368-75; September, 1926.
200. RUCH, G. M., and STODDARD, GEORGE D. *Correlation Chart.* Iowa City, Ia.: University Book Store (2-4 South Clinton Street).
201. RUCH, G. M.; ACKERSON, LUTTON; and JACKSON, JESSE D. "An Empirical Study of the Spearman-Brown Formula as Applied to Educational Test Material." *Journal of Educational Psychology* 17: 309-13; May, 1926.
202. RUCH, G. M. "Minimum Essentials in Reporting Data on Standard Tests." *Journal of Educational Research* 12: 349-58; December, 1925.
203. RUCH, G. M., and OTHERS. *Objective Examination Methods in the Social Studies.* Chicago: Scott, Foresman and Co., 1926. p. 105-16.
204. RUCH, G. M. *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Co., 1929. 478 p.
205. RUCH, G. M., and KOERTH, W. " 'Power' vs. 'Speed' in Army Alpha." *Journal of Educational Psychology* 14: 193-208; April, 1923.
206. RUCH, G. M. "The Speed Factor in Mental Measurements." *Journal of Educational Research* 9: 39-45; January, 1924.
207. RUCH, G. M., and STODDARD, GEORGE D. *Tests and Measurements in High School Instruction.* Yonkers-on-Hudson, N. Y.: World Book Co., 1927. 381 p.
208. RUGER, H. A. *Correlation Chart.* New York: Teachers College, Columbia University.
209. RUGG, HAROLD O. *Statistical Methods Applied to Education.* Boston: Houghton Mifflin Co., 1917. 410 p.
210. RULON, PHILLIP J. "A Graph for Estimating Reliability in One Range, Knowing It in Another." *Journal of Educational Psychology* 21: 140-42; February, 1930.
211. SALISBURY, FRANK S. "A Simplified Method of Computing Multiple Correlation Constants." *Journal of Educational Psychology* 20: 44-52; January, 1929.
212. SCATES, DOUGLAS E., and NOFFSINGER, FOREST R. "Factors Which Determine the Effectiveness of Weighting." *Journal of Educational Research* 24: 280-85; November, 1931.
213. SHEN, EUGENE. "A Note on the Standard Error of the Spearman-Brown Formula." *Journal of Educational Psychology* 17: 93-94; February, 1926.
214. SHEN, EUGENE. "The Standard Error of Certain Estimated Coefficients of Correlation." *Journal of Educational Psychology* 15: 462-65; October, 1924.
215. SLOCOMBE, CHARLES S. "A Further Note on the Use of the Spearman Prophecy Formula: A Correction." *Journal of Educational Psychology* 18: 347-48; May, 1927.
216. SLOCOMBE, CHARLES S. "The Spearman Prophecy Formula." *Journal of Educational Psychology* 18: 125-26; February, 1927.
217. SMITH, HERBERT E. *The Validity and Reliability of Teachers' Judgments of Difficulty in Curricular Material.* Unpublished doctor's thesis, University of California, 1929. 84 p.

218. SMITH, HERBERT E. "The Validity of Teachers' Judgments of Difficulty of Curricular Material." *Journal of Educational Psychology* 21: 460-66; September, 1930.
219. SPEARMAN, C. "The Proof and Measurement of Association between Two Things." *American Journal of Psychology* 15: 72-101; January, 1904.
220. STAFFELBACH, ELMER H. "Weighting Responses in True-False Examinations." *Journal of Educational Psychology* 21: 136-39; February, 1930.
221. STARCH, DANIEL. *Educational Measurements.* New York: Macmillan Co., 1916. 202 p.
222. STEBBINS, RENA, and PECHSTEIN, L. A. "Quotients I, E, and A." *Journal of Educational Psychology* 13: 385-98; October, 1922.
223. SYMONDS, PERCIVAL M. "The Accuracy of Certain Standard Tests for School Sectioning and Marking." *Journal of Educational Psychology* 15: 423-32; October, 1924.
224. SYMONDS, PERCIVAL M. "Choice of Items for a Test on the Basis of Difficulty." *Journal of Educational Psychology* 20: 481-93; October, 1929.
225. SYMONDS, PERCIVAL M. "Factors Influencing Test Reliability." *Journal of Educational Psychology* 19: 73-87; February, 1928.
226. SYMONDS, PERCIVAL M. *Measurement in Secondary Education.* New York: Macmillan Co., 1927. 588 p.
227. SYMONDS, PERCIVAL M. "Variations of the Product-Moment (Pearson) Coefficient of Correlation." *Journal of Educational Psychology* 17: 458-69; October, 1926.
228. TALBOTT, E. O., and RUCH, G. M. "Minor Studies on Objective Examination Methods II. The Theory of Sampling as Applied to Examinations." *Journal of Educational Research* 20: 199-206; October, 1929.
229. TERMAN, LEWIS M. *The Measurement of Intelligence.* Boston: Houghton Mifflin Co., 1916. 362 p.
230. THOMSON, GODFREY H., and PINTNER, RUDOLF. "Spurious Correlations and Relationship between Tests." *Journal of Educational Psychology* 15: 433-44; October, 1924.
231. THORNDIKE, EDWARD L. *An Introduction to the Theory of Mental and Social Measurements.* 2d ed. rev. and enl. New York: Teachers College, Columbia University, 1913. 277 p.
232. THURSTONE, L. L. "The Absolute Zero in Intelligence Measurement." *Psychological Review* 35: 175-97; May, 1928.
233. THURSTONE, L. L. *Correlation Chart.* Chicago: C. H. Stoelting Co. (3037 Carroll Avenue).
234. THURSTONE, L. L. "Equally Often Noticed Differences." *Journal of Educational Psychology* 18: 289-93; May, 1927.
235. THURSTONE, L. L. "A Method of Scaling Psychological and Educational Tests." *Journal of Educational Psychology* 16: 433-51; October, 1925.
236. THURSTONE, L. L. "A Note on the Spearman-Brown Formula." *Journal of Experimental Psychology* 11: 62-63; February, 1928.
237. THURSTONE, L. L. "Scale Construction with Weighted Observations." *Journal of Educational Psychology* 19: 441-53; October, 1928.
238. THURSTONE, L. L. "A Scoring Method for Mental Tests." *Psychological Bulletin* 16: 235-40; July, 1919.
239. THURSTONE, L. L. "The Scoring of Individual Performance." *Journal of Educational Psychology* 17: 446-57; October, 1926.
240. THURSTONE, L. L. "The Unit of Measurement in Educational Scales." *Journal of Educational Psychology* 18: 505-24; November, 1927.
241. THURSTONE, THELMA G. "The Difficulty of a Test and Its Diagnostic Value." *Journal of Educational Psychology* 23: 335-43; May, 1932.
242. TOLLEY, H. R., and EZEKIEL, MORDECAI. "The Doolittle Method for Solving Multiple Correlation Equations versus the Kelley-Salisbury 'Iteration' Method." *Journal of the American Statistical Association* 22: 497-500; December, 1927.
243. TOLLEY, H. R., and EZEKIEL, MORDECAI. "A Method of Handling Multiple Correlation Problems." *Journal of the American Statistical Association* 18: 993-1003; December, 1923.
244. TOOPS, HERBERT A., and EDGERTON, HAROLD A. "An Abac for Determining the Probable Correlation over a Larger Range Knowing It over a Shorter One." *Journal of Educational Research* 16: 382-85; December, 1927.

245. TOOPS, HERBERT A. *Correlation Chart.* Columbus, O.: the Author (Ohio State University).
246. TOOPS, HERBERT A. "Two Devices for Aiding Calculation." *Journal of Experimental Psychology* 9: 60-66; February, 1926.
247. TOOPS, HERBERT A., and SYMONDS, PERCIVAL M. "What Shall We Expect of the AQ?" *Journal of Educational Psychology* 13: 513-28, December, 1922; 14: 27-38, January, 1923.
248. TRABUE, MARION R. *Completion-Test Language Scales.* Contributions to Education, No. 77. New York: Teachers College, Columbia University, 1916. 118 p.
249. VAN WAGENEN, MARVIN J. "Some Implications of the Revised Van Wagenen History Scales." *Teachers College Record* 27: 142-48; October, 1925.
250. VINCENT, LEONA. *A Study of Intelligence Test Elements.* Contributions to Education, No. 152. New York: Teachers College, Columbia University, 1924. 36 p.
251. WARREN, RICHARD, and MENDENHALL, ROBERT M. *The Mendenhall-Warren-Hollerith Correlation Method.* Document No. 1. New York: Statistical Bureau, Columbia University, 1929.
252. WEIDEMANN, CHARLES C. *How to Construct the True-False Examination.* Contributions to Education, No. 225. New York: Teachers College, Columbia University, 1926. 118 p.
253. WEIDEMANN, CHARLES C. "The 'Omission' as a Specific Determiner in the True-False Examination." *Journal of Educational Psychology* 22: 435-39; September, 1931.
254. WEST, PAUL V. "The Significance of Weighted Scores." *Journal of Educational Psychology* 15: 302-8; May, 1924.
255. WESTON, S. BURNS, and ENGLISH, HORACE B. "The Influence of the Group on Psychological Test Scores." *American Journal of Psychology* 37: 600-1; October, 1926.
256. WHELDEN, CHESTER H., and DAVIES, F. J. J. "A Method for Judging the Discrimination of Individual Questions on True-False Examinations." *Journal of Educational Psychology* 22: 290-306; April, 1931.
257. WHIPPLE, GUY M. *Manual of Mental and Physical Tests.* 2d rev. and enl. ed. Baltimore: Warwick and York, 1914-15. 2 vols.
258. WILLSON, G. M. "Standard Deviations versus Age as a Score Unit." *Journal of Educational Research* 13: 189-96; March, 1926.
259. WILSON, WILLIAM R. "The Misleading Accomplishment Quotient." *Journal of Educational Research* 17:1-10; January, 1928.
260. WOOD, BEN D. *Measurement in Higher Education.* Yonkers-on-Hudson, N. Y.: World Book Co., 1923. 337 p.
261. WOOD, BEN D. "Studies of Achievement Tests." *Journal of Educational Psychology* 17: 1-22; 125-39; 263-69; January, February, and April, 1926.
262. WOOD, ELEANOR PERRY. "Improving the Validity of Collegiate Achievement Tests." *Journal of Educational Psychology* 18: 18-25; January, 1927.
263. WOODWORTH, ROBERT S. "Combining the Results of Several Tests." *Psychological Review* 19: 97-123; March, 1912.
264. YERKES, ROBERT M., editor. *Psychological Examining in the United States Army.* Memoirs of the National Academy of Sciences, Vol. 15. Washington, D. C.: Government Printing Office, 1921. 90 p.

Chapter IV. Recent Developments in Testing for Guidance

265. ANDERSON, VICTOR VANCE. *Psychiatry in Industry.* New York: Harper and Brothers, 1929. 364 p.
266. BINGHAM, WALTER VAN DYKE, and FREYD, MAX. *Procedures in Employment Psychology.* Chicago: A. W. Shaw Co., 1926. 269 p.
267. BROTEMARKLE, R. A. "Clinical Psychology and Student Personnel Work." *Personnel Journal* 10: 254-58; December, 1931.
268. COWDERY, KARL M. "Measurement of Professional Attitudes; Differences between Lawyers, Physicians, and Engineers." *Journal of Personnel Research* 5: 131-41; August, 1926.

269. FRYER, DOUGLAS. *Measurement of Interests in Relation to Human Adjustment.* New York: Henry Holt and Co., 1931. 488 p.
270. HAMMOND, H. P., and STODDARD, GEORGE D. *A Study of Placement Examinations.* University of Iowa Studies in Education, Vol. 4, No. 7. Iowa City: the University, 1928. 59 p.
271. HARTSHORNE, HUGH, and MAY, MARK A. *Studies in Deceit.* New York: Macmillan Co., 1928. Book I, 414 p.; Book II, 306 p.
272. HELLER, WALTER T. "Industrial Psychology and Its Development in Switzerland." *Personnel Journal* 8: 435-41; April, 1930.
273. HULL, CLARK L. *Aptitude Testing.* Yonkers-on-Hudson, N. Y.: World Book Co., 1928. 536 p.
274. KITSON, HARRY D. *The Psychology of Vocational Adjustment.* Philadelphia: J. B. Lippincott Co., 1925. 274 p.
275. KELLEY, TRUMAN LEE. *Crossroads in the Mind of Man.* Stanford University, Calif.: Stanford University Press, 1928. 238 p.
276. KELLEY, TRUMAN LEE. *Interpretation of Educational Measurements.* Yonkers-on-Hudson, N. Y.: World Book Co., 1927. 363 p.
277. KELLEY, TRUMAN LEE, and SALISBURY, FRANK S. "An Iteration Method for Determining Multiple Correlation Constants." *Journal of the American Statistical Association* 21: 282-92; September, 1926.
278. KORNHAUSER, ARTHUR W. "Industrial Psychology in England, Germany, and the United States." *Personnel Journal* 8: 421-34; April, 1930.
279. KORNHAUSER, ARTHUR W., and KINGSBURY, F. A. *Psychological Tests in Business.* Chicago: University of Chicago Press, 1924. 194 p.
280. LAWRENCE, JAMES C., and OTHERS. "Unemployment Educational and Guidance Problems." *Journal of Adult Education* 4: 266-78; June, 1932.
281. LINK, HENRY C. *Employment Psychology.* New York: Macmillan Co., 1919. 440 p.
282. LINK, HENRY C. "Vocational Guidance in Practice." A chapter from *Parent and Child,* Sidonie Gruenberg, editor, to be published by the Child Study Association of America.
283. MILLER, LAWRENCE W. *An Experimental Study of the Iowa Placement Examinations.* University of Iowa Studies in Education, Vol. 5, No. 6. Iowa City: the University, 1930. 116 p.
284. MOSS, FRED A. "Preliminary Report on Medical Aptitude Tests for 1931-32." *School and Society* 34: 132-34; July 25, 1931.
285. MYERS, GEORGE E. *The Problem of Vocational Guidance.* New York: Macmillan Co., 1927. 311 p.
286. O'CONNOR, JOHNSON. *Born That Way.* Baltimore: Williams and Wilkins Co., 1928. 323 p.
287. PATERSON, DONALD G., and OTHERS. *Minnesota Mechanical Ability Tests.* Minneapolis: University of Minnesota Press, 1930. 586 p.
288. PATERSON, DONALD G. "The Minnesota Unemployment Research Project." *Personnel Journal* 10: 318-28; February, 1932.
289. PATERSON, DONALD G. *Physique and Intellect.* New York: Century Co., 1930. 304 p.
290. PINTNER, RUDOLF. *Intelligence Testing.* New York: Henry Holt and Co., 1931. 555 p.
291. POND, MILLICENT. "What Is New in Employment Testing." *Personnel Journal* 11: 10-16; June, 1932.
292. PROCTOR, WILLIAM M. *Educational and Vocational Guidance.* Boston: Houghton Mifflin Co., 1925. 352 p.
293. REMMERS, H. H. *Achievement of Our High Schools; Results of the State High School Testing Program, 1930-1931.* Purdue University Bulletin, Vol. 32, No. 2, Studies in Higher Education, No. 18. Lafayette, Ind.: the University, 1931. 30 p.
294. SEASHORE, CARL E. "The Individual in Mass Education." *School and Society* 23: 569-76; May 8, 1926.
295. SEASHORE, CARL E. *The Psychology of Musical Talent.* New York: Silver, Burdett and Co., 1919. 288 p.
296. SEASHORE, CARL E. "Recognition of the Individual." *Bulletin of Engineering Education* 15: 1-12.

297. Spearman, Charles E. *The Abilities of Man.* New York: Macmillan Co., 1927. 415 p.
298. Stanton, Hazel M. *Psychological Tests of Musical Talent.* Rochester, N. Y.: Eastman School of Music, University of Rochester, 1925. 48 p.
299. Stenquist, John L. *Measurements of Mechanical Ability.* Contributions to Education, No. 130. New York: Teachers College, Columbia University, 1923. 101 p.
300. Stoddard, George D. *Iowa Placement Examinations.* University of Iowa Studies in Education, Vol. 3, No. 2. Iowa City: the University, 1925. 103 p.
301. Stoddard, George D. "Iowa Placement Examinations." *University of Iowa Studies in Psychology.* Psychological Monographs, Vol. 39. Princeton, N. J.: Psychological Review Co., 1928. p. 92-101.
302. Strong, Edward K. *Change of Interests with Age.* Stanford University, Calif.: Stanford University Press, 1931. 235 p.
303. Strong, Edward K. "Diagnostic Value of the Vocational Interest Test." *Educational Record* 10: 59-68; January, 1929.
304. Stuart, Milo H., and Morgan, DeWitt S. *Guidance at Work.* New York: McGraw-Hill Book Co., 1931. 104 p.
305. Symonds, Percival M. *Diagnosing Personality and Conduct.* New York: Century Co., 1931. 602 p.
306. Symonds, Percival M. *Tests and Interest Questionnaires in the Guidance of High School Boys.* New York: Teachers College, Columbia University, 1930. 61 p.
307. Thurstone, L. L., and Chave, E. J. *The Measurement of Attitude.* Chicago: University of Chicago Press, 1929. 96 p.
308. Trabue, Marion R. "Occupational Ability Patterns." *Personnel Journal* (in press).
309. Unger, Edna W., and Burr, Emily T. *Minimum Mental Age Levels of Accomplishment.* Albany, N. Y.: University of the State of New York Press, 1931. 108 p.
310. Viteles, Morris S. *Industrial Psychology.* New York: W. W. Norton and Co., 1932. 652 p.
311. Viteles, Morris S. "Vocational Guidance of Adults." *Personnel Journal* 10: 335-41; February, 1932.
312. Waples, Douglas, and Tyler, Ralph W. *What People Want to Read About.* Chicago: University of Chicago Press, 1931. 312 p.
313. White House Conference on Child Health and Protection. *Vocational Guidance.* New York: Century Co., 1932. 396 p.
314. Wright, Sewall. "The Theory of Path Coefficients." *Genetics* 8: 238-55.
315. Yoakum, Clarence S., and Yerkes, R. M. *Army Mental Tests.* New York: Henry Holt and Co., 1920. 303 p.

Chapter V. Recent Developments in the Use of Tests

316. Adams, Mary A. "City-wide Experimentation in the Intermediate Curriculum Program in Geography and History." *Baltimore Bulletin of Education* 9: 172-77; April, 1931.
317. Bamesberger, Velda C. *An Appraisal of a Social Studies Course.* Contributions to Education, No. 328. New York: Teachers College, Columbia University, 1928. 91 p.
318. Barr, A. S. *An Introduction to the Scientific Study of Classroom Supervision.* New York: D. Appleton and Co., 1931. p. 65-97.
319. Billett, Roy O. "What the High Schools Are Doing for the Individual." *Bulletin (Proceedings)* No. 40: 139-68; March, 1932. Berwyn, Ill.: Department of Secondary School Principals, National Education Association.
320. Book, William F., and Norvell, Lee. "The Will to Learn." *Pedagogical Seminary* 29: 305-62; December, 1922.
321. Bowman, Herbert L. *The Relation of Reported Preference to Performance in Problem Solving.* University of Missouri Bulletin, Vol. 30, Education Series, No. 29. Columbia, Mo.: the University, 1929. 52 p.
322. Brigham, Carl C. *A Study of Error.* New York: College Entrance Examination Board, 1932. 384 p.

323. BROENING, ANGELA MARIE. "Co-operative Curriculum Research in English." *Baltimore Bulletin of Education* 10: 10-13; September, 1931.
324. BROENING, ANGELA MARIE. *Developing Appreciation through Teaching Literature.* Johns Hopkins University Studies in Education, No. 13. Baltimore: Johns Hopkins University Press, 1929. 118 p.
325. BROENING, GRACE D. *An Evaluation of Individualized Instruction in Junior High Schools Experimentally Determined.* Unpublished master's thesis, Johns Hopkins University, 1926.
326. BRUECKNER, LEO J., and OTHERS. *Diagnostic Tests and Practice Exercises in Arithmetic.* Philadelphia: John C. Winston Co., 1928.
327. BUCKINGHAM, B. R. "Individualizing Instruction on the Basis of Testing, with Special Reference to Arithmetic." *Proceedings of Second Annual Conference on Educational Research and Guidance.* San Jose Teachers College Bulletin. Sacramento: California State Printing Office, 1923. p. 16-32.
328. BURCH, MARY CROWELL. *Determination of a Content of the Course in Literature of a Suitable Difficulty for Junior and Senior High School Students.* Genetic Psychology Monographs, Vol. 4, Nos. 2-3. Worcester, Mass.: Clark University, 1928. p. 69-332.
329. BURR, MARVIN Y. *A Study of Homogeneous Grouping.* Contributions to Education, No. 457. New York: Teachers College, Columbia University, 1931. 69 p.
330. BURTON, W. H. "Evaluation of Supervision through Precise Measurement." *Educational Supervision.* First Yearbook. National Conference on Educational Method. New York: Teachers College, Columbia University, 1928. p. 220-26.
331. BUSWELL, GUY T., and JUDD, CHARLES H. *Summary of Educational Investigations Relating to Arithmetic.* Supplementary Educational Monographs, No. 27. Chicago: University of Chicago Press, 1925. 212 p.
332. BUSWELL, GUY T. Supplements entitled "Summary of Arithmetic Investigations," appearing annually in the *Elementary School Journal.*
333. CASWELL, HOLLIS LELAND. *City School Surveys.* Contributions to Education, No. 358. New York: Teachers College, Columbia University, 1929. 130 p.
334. CATTELL, J. MCKEEN. "Mental Tests and Measurements." *Mind* 15: 373-81; July, 1890.
335. CHISM, LESLIE L. "Classification and Promotion Practices in the Elementary School." *Elementary School Journal* 33: 89-91; October, 1932.
336. CLAPP, FRANK L., and YOUNG, ROBERT V. *Clapp-Young Self-Marking Tests.* New York: Houghton Mifflin Co., 1932.
337. COLLIER, PAUL D., and MILLER, HARLAN H. "What Types of Achievement Tests Should Be Used as a Basis of Promotion?" *Junior High Clearing House* 3: 45-47; April-May, 1929.
338. COLLINGS, ELLSWORTH. *An Experiment with a Project Curriculum.* New York: Macmillan Co., 1923. p. 9, 233-59.
339. CORYELL, NANCY GILLMORE. *An Evaluation of Extensive and Intensive Teaching of Literature.* Contributions to Education, No. 275. New York: Teachers College, Columbia University, 1927. 201 p.
340. COURTIS, STUART APPLETON. *Courtis Standard Practice Tests in Arithmetic.* Yonkers-on-Hudson, N. Y.: World Book Co., 1920.
341. COURTIS, STUART APPLETON. *Why Children Succeed.* Detroit: Courtis Standard Tests (1807 E. Grand Blvd.), 1925. 271 p.
342. COY, GENEVIEVE L. "A Study of Various Factors Which Influence the Use of the Accomplishment Quotient as a Measure of Teaching Efficiency." *Journal of Educational Research* 21: 29-42; January, 1930.
343. CURTIS, FRANCIS D., and WOODS, GERALD G. "A Study of the Relative Teaching Values of Four Common Practices in Correcting Examination Papers." *School Review* 37: 615-23; October, 1929.
344. CUTRIGHT, PRUDENCE. "The Use of Research in Supervision and Curriculum Construction." *Educational Supervision.* First Yearbook. National Conference on Educational Method. New York: Teachers College, Columbia University, 1928. Chapter 12, p. 157-75.
345. DOUGLASS, CARLETON E. "Promotional Standards in the Intermediate Grades." *Baltimore Bulletin of Education* 9: 86-87; December, 1930.

346. Dunlap, Jack Wilbur, and Kurtz, Albert K. *Handbook of Statistical Nomographs, Tables, and Formulas.* Yonkers-on-Hudson, N. Y.: World Book Co., 1932. 163 p.
347. Dvorak, August, and English, Elsie. "The Efficiency of Remedial Teaching." *Educational Administration and Supervision* 18: 466-71; September, 1932.
348. Edgerton, Harold A. "A Graphic Method of Finding Standard Errors and Probable Errors of Differences." *Journal of Educational Psychology* 23: 56-57; January, 1932.
349. *Education Index.* Edited by Isabel L. Towner and Dorothy R. Carpenter. New York: H. W. Wilson Co., 1932. January, 1929, to June, 1932. 1891 p.
350. Eells, Walter Crosby. "Use of Standard Tests in 72 Published School Surveys." *School Life* 14: 168-69; May, 1929.
351. Field, Helen A. *Extensive Individual Reading Versus Class Reading.* Contributions to Education, No. 394. New York: Teachers College, Columbia University, 1930. 52 p.
352. Frazee, Laura. "Standards of Promotion for Primary Grades." *Baltimore Bulletin of Education* 9: 79-85; December, 1930.
353. Galton, Francis. *Hereditary Genius: An Inquiry Into Its Laws and Consequences.* New and rev. ed. New York: D. Appleton and Co., 1871. 390 p.
354. Garrett, Henry E. *Statistics in Psychology and Education.* New York: Longmans, Green and Co., 1926. 317 p.
355. Gates, Arthur I. "Characteristics and Uses of Practice Exercises in Reading." *Teachers College Record* 32: 221-35; December, 1930.
356. Gates, Arthur I. "Diagnosis and Measurement of Reading Ability by Intrinsic Methods." *Modern Education* 3: 2-3, 42-46; April, 1931.
357. Gates, Arthur I. *The Improvement of Reading.* New York: Macmillan Co., 1927. 440 p.
358. Gates, Arthur I. *Interest and Ability in Reading.* New York: Macmillan Co., 1931. 264 p.
359. Gilliland, Adam Raymond; Jordan, R. H.; and Freeman, Frank S. *Educational Measurements and the Classroom Teacher.* New York: Century Co., 1931. 400 p.
360. Goodenough, Florence L. "Efficiency in Learning and the Accomplishment Ratio." *Journal of Educational Research* 12: 297-300; November, 1925.
361. Gray, William S. *Summary of Investigations Relating to Reading.* Supplementary Educational Monographs, No. 28. Chicago: University of Chicago Press, 1925. 275 p.
362. Gray, William S. Supplements entitled "Summary of Investigations Relating to Reading" appearing annually in the *Elementary School Journal.*
363. Greene, Harry A., and Jorgensen, A. N. *The Use and Interpretation of Educational Tests.* New York: Longmans, Green and Co., 1929. 389 p.
364. Gregg, Russell T., and Hamilton, Thomas T. *Annotated Bibliography of Graduate Theses in Education at the University of Illinois.* University of Illinois Bulletin, Vol. 28, No. 40, Educational Research Bulletin, No. 55. Urbana, Ill.: the University, 1931. 80 p.
365. Guiler, W. S. *The Ohio Survey of English Usage.* Report based on the results of the Every Pupil Test administered by the State Department of Education, December 2, 1930. Columbus: Ohio State Department of Education, 1931.
366. Hanna, Paul R. *Arithmetic Problem Solving.* New York: Teachers College, Columbia University, 1929. 68 p.
367. Harap, Henry. *Technique of Curriculum Making.* New York: Macmillan Co., 1928. p. 215-21.
368. Hildreth, Gertrude H. *Psychological Service for School Problems.* Yonkers-on-Hudson, N. Y.: World Book Co., 1930. 317 p.
369. Hollingshead, Arthur D. *An Evaluation of the Use of Certain Educational and Mental Measurements for Purposes of Classification.* Contributions to Education, No. 302. New York: Teachers College, Columbia University, 1928. 63 p.
370. Holroyd, Flora E. *A Supervisory Project in Educational Measurements.* Kansas State Teachers College Educational Monograph, No. 1. Pittsburg, Kans.: the College, 1932. 72 p.
371. Holzinger, Karl J. *Statistical Methods for Students in Education.* Boston: Ginn and Co., 1928. 372 p.

372. HULL, CLARK L. *Aptitude Testing.* Yonkers-on-Hudson, N. Y.: World Book Co., 1928. 536 p.
373. HULL, CLARK L. "An Automatic Machine for Making Multiple Aptitude Forecasts." *Journal of Educational Psychology* 16: 593-98; December, 1925.
374. HYDLE, L. L., and CLAPP, FRANK L. *Elements of Difficulty in the Interpretation of Concrete Problems in Arithmetic.* University of Wisconsin, Bureau of Educational Research Bulletin, No. 9. Madison, Wis.: the University, 1927. 84 p.
375. IRION, THEODORE W. H. *Comprehension Difficulties of Ninth Grade Students in the Study of Literature.* Contributions to Education, No. 189. New York: Teachers College, Columbia University, 1925. 116 p.
376. KELIHER, ALICE V. *A Critical Study of Homogeneous Grouping in Elementary Schools.* Contributions to Education, No. 452. New York: Teachers College, Columbia University, 1930. 165 p.
377. KELLEY, TRUMAN LEE. *Correlation Chart.* Yonkers-on-Hudson, N. Y.: World Book Co., 1927.
378. KELLEY, TRUMAN LEE. *Interpretation of Educational Measurements.* Yonkers-on-Hudson, N. Y.: World Book Co., 1927. 363 p.
379. KNIGHT, F. B. "Possibilities of Objective Techniques in Supervision." *Journal of Educational Research* 16: 1-15; June, 1927.
380. KRAMER, GRACE A. *The Effect of Certain Factors in the Verbal Arithmetic Problems on Children's Success in the Solution.* Baltimore: Johns Hopkins University. (In Press.)
381. KRAMER, GRACE A. "The Relation of the Testing Program to Promotion." *Baltimore Bulletin of Education* 9: 73-78; December, 1930.
382. LEONARD, JOHN PAUL. *The Use of Practice Exercises in the Teaching of Capitalization and Punctuation.* Contributions to Education, No. 372. New York: Teachers College, Columbia University, 1930. 78 p.
383. LINCOLN, EDWARD A. *Beginnings in Educational Measurement.* Philadelphia: J. B. Lippincott Co., 1924. 151 p.
384. LYMAN, ROLLO L. *Summary of Investigations Relating to Grammar, Language, and Composition.* Supplementary Educational Monographs, No. 36. Chicago: University of Chicago, 1929. 302 p.
385. MACDONALD, MARION E. *Practical Statistics for Teachers.* New York: Macmillan Co., 1930. 176 p.
386. MADSEN, I. N. *Educational Measurement in the Elementary Grades.* Yonkers-on-Hudson, N. Y.: World Book Co., 1930. Chapter 10, "Educational Uses of Standardized Tests," p. 220-67.
387. MANN, C. V. "Objective Type Tests in Engineering Drawing and Descriptive Geometry; Summary of Research in Drawing and Descriptive Geometry at Missouri School of Mines, for the Year 1927-28." *Journal of Engineering Education* 19: 979-92: June, 1929.
388. MCCALL, WILLIAM A. *How to Measure in Education.* New York: Macmillan Co., 1922. 416 p.
389. METCALF, ARTHUR ANSEL. "Diagnostic Testing and Remedial Teaching." *School Executives Magazine* 49: 358-60; April, 1930.
390. MICHELL, ELENE. *Teaching Values in New-Type History Tests.* Yonkers-on-Hudson, N. Y.: World Book Co., 1930. 179 p.
391. MONROE, MARION. *Methods for Diagnosis and Treatment of Cases of Reading Disability.* Genetic Psychology Monographs, Vol. 4, Nos. 4-5. Worcester, Mass.: Clark University, 1928. p. 335-456.
392. MONROE, WALTER S., and ENGELHART, MAX D. *A Critical Summary of Research Relating to the Teaching of Arithmetic.* University of Illinois Bulletin, Vol. 29, No. 5, Educational Research Bulletin, No. 58. Urbana, Ill.: the University, 1931. 115 p.
393. MONROE, WALTER S. *How Pupils Solve Problems in Arithmetic.* University of Illinois Bulletin, Vol. 26, No. 23, Educational Research Bulletin, No. 44. Urbana, Ill.: the University, 1929. 31 p.
394. MONROE, WALTER S., and OTHERS. *Locating Educational Information in Published Sources.* University of Illinois Bulletin, Vol. 27, No. 45, Educational Research Bulletin, No. 50. Urbana, Ill.: the University, 1930. 142 p.

395. MONROE, WALTER S. *Measuring the Results of Teaching.* Boston: Houghton Mifflin Co., 1918. 297 p.
396. MONROE, WALTER S., and OTHERS. *Ten Years of Educational Research, 1918-1927.* University of Illinois Bulletin, Vol. 25, No. 51, Educational Research Bulletin, No. 42. Urbana, Ill.: the University, 1928. 367 p.
397. MORTON, R. L. "An Analysis of Errors in the Solution of Arithmetic Problems." *Educational Research Bulletin (Ohio State University)* 4: 187-90; April 29, 1925.
398. MORTON, R. L. "Pupils' Errors in Solving Arithmetic Problems." *Educational Research Bulletin (Ohio State University)* 4: 155-58; April 15, 1925.
399. MORTON, R. L. "Solving Arithmetic Problems: Case Studies." *Educational Research Bulletin (Ohio State University)* 4: 203-7; May 13, 1925.
400. NATIONAL EDUCATION ASSOCIATION, DEPARTMENT OF SUPERINTENDENCE. "Character Education." *Tenth Yearbook.* Washington, D. C.: the Association, 1932. Chapter 16, "Tests and Measurement in Character Education," p. 345-404.
401. NATIONAL EDUCATION ASSOCIATION, DEPARTMENT OF SUPERINTENDENCE. "Development of High School Curriculum." *Sixth Yearbook.* Washington, D. C.: the Association, 1928. 584 p.
402. NATIONAL EDUCATION ASSOCIATION, DEPARTMENT OF SUPERINTENDENCE. "Five Unifying Factors in American Education." *Ninth Yearbook.* Washington, D. C.: the Association, 1931. p. 23-79, 115, 118, 120-25.
403. NATIONAL SOCIETY FOR THE STUDY OF EDUCATION. *Adapting the Schools to Individual Differences.* Twenty-Fourth Yearbook, Part 2. Bloomington, Ill.: Public School Publishing Co., 1925. 410 p.
404. NEWLUN, CHESTER O. *Teaching Children to Summarize in Fifth Grade History.* Contributions to Education, No. 404. New York: Teachers College, Columbia University, 1930. 75 p.
405. OBRIEN, FRANCIS P. "Supervisory Assistance in Teaching Geography and History." *University of Kansas Bulletin of Education* 3: 1-32; February, 1932.
406. ODELL, C. W. *Educational Statistics.* New York: Century Co., 1925. 334 p.
407. ORLEANS, JACOB S., and SEALY, GLENN A. *Objective Tests.* Yonkers-on-Hudson, N. Y.: World Book Co., 1928. Chapter 6, p. 87-96.
408. OSBURN, WORTH J. *Corrective Arithmetic for Supervisors, Teachers, and Teacher-Training Classes.* Boston: Houghton Mifflin Co., 1924-1929. 2 vols.
409. O'SHEA, HARRIET EASTABROOKS. *A Study of the Effect of the Interest of a Passage on Learning Vocabulary.* Contributions to Education, No. 351. New York: Teachers College, Columbia University, 1930. 122 p.
410. OTIS, ARTHUR S. *Correlation Chart.* Yonkers-on-Hudson, N. Y.: World Book Co., 1922.
411. OTIS, ARTHUR S. *Statistical Method in Educational Measurement.* Yonkers-on-Hudson, N. Y.: World Book Co., 1925. Chapter 21, "Grading and Classifying," p. 267-88.
412. OTIS, ARTHUR S. *Universal Percentile Graph.* Yonkers-on-Hudson, N. Y.: World Book Co., 1924. 4 p.
413. PETERS, CHARLES C., and WYKES, ELIZABETH CROSSLEY. "Simplified Methods for Computing Regression Coefficients and Partial and Multiple Correlations." *Journal of Educational Research* 23: 383-93; May, 1931.
414. PRESSEY, SIDNEY L., and PRESSEY, LUELLA COLE. *Introduction to the Use of Standard Tests.* rev. ed. Yonkers-on-Hudson, N. Y.: World Book Co., 1931. 266 p.
415. PRESSEY, SIDNEY L. "A Statistical Study of Children's Errors in Sentence Structure." *English Journal* 14: 529-35; September, 1925.
416. PURDOM, T. LUTHER. *The Value of Homogeneous Grouping.* Baltimore: Warwick and York, 1929. 99 p.
417. RAGUSE, FLORENCE W. "Qualitative and Quantitative Achievements in First Grade Reading." *Teachers College Record* 32: 424-36; February, 1931.
418. RANKIN, PAUL T. "Survey Techniques for the Experimental Determination of the Value of Materials and Methods." *Scientific Method in Supervision.* Second Yearbook. National Conference of Supervisors and Directors of Instruction. New York: Teachers College, Columbia University, 1929. Chapter 17, p. 223-34.
419. RENWICK, E. M. "Children's Difficulties in the Study of Mensuration." *Forum of Education* 7: 106-20; June, 1929.

420. ROLKER, EDNA. "After Testing in Baltimore, What?" *Baltimore Bulletin of Education* 7: 36-38; November, 1928.
421. ROLKER, EDNA. "The Spread of Ability in Arithmetic and Its Relation to Standards of Promotion and Revision of the Course of Study in Grades Four, Five and Six." *Baltimore Bulletin of Education* 9: 5-7; September, 1930.
422. RUCH, G. M., and OTHERS. *Compass Diagnostic Tests in Arithmetic.* New York: Scott, Foresman and Co., 1925.
423. RUCH, G. M. *The Objective or New-Type Examination.* New York: Scott, Foresman and Co., 1929. 478 p.
424. RUCH, G. M. "Recent Experiments on New-Type Examinations." *Los Angeles Educational Research Bulletin* 10: 2-5, 8; March, 1930.
425. RUCH, G. M., and STODDARD, G. D. *Tests and Measurements in High School.* Yonkers-on-Hudson, N. Y.: World Book Co., 1927. 381 p.
426. RUSSELL, CHARLES. *Standard Tests.* Boston: Ginn and Co., 1930. Chapters 13-14, p. 304-409.
427. SANGREN, PAUL V. *Improvement of Reading through the Use of Tests.* Kalamazoo, Mich.: Western State Teachers College, 1932. 207 p.
428. SEATON, J. T., and PRESSEY, S. L. *Minimal Essentials Tests in English Composition.* Dover, Del.: State Department of Public Instruction, 1931.
429. SMITH, DORA V. *Class Size in High School English.* Minneapolis: University of Minnesota Press, 1931. 309 p.
430. SMITH, HENRY LESTER, and WRIGHT, WENDELL WILLIAM. *Tests and Measurements.* New York: Silver, Burdett and Co., 1928. 540 p.
431. SMITH, HOMER J. "Objective Measurement in Industrial Education." *Industrial Education Magazine* 31: 331-36; March, 1930.
432. SPENCER, PETER L. "The Improvement of Teaching by Means of 'Homemade' Non-Standard Diagnostic Tests and Remedial Instruction." *School Review* 31: 276-81; April, 1923.
433. STENQUIST, JOHN L. "Baltimore Constantly Checks Results." *Journal of Education* 112: 183-85; September 22, 1930.
434. STENQUIST, JOHN L. *Teachers Class Analysis Charts.* Baltimore: Bureau of Edutional Research, Department of Education, 1927-1932.
435. STREITZ, RUTH. *Teachers' Difficulties in Arithmetic and Their Correctives.* University of Illinois Bulletin, Vol. 21, No. 34, Educational Research Bulletin, No. 18. Urbana, Ill.; the University, 1924. 34 p.
436. SYMONDS, PERCIVAL M., and CHASE, DORIS HARTER. "Practice vs. Motivation." *Journal of Educational Psychology* 20: 19-35; January, 1929.
437. TAYLOR, JOHN CAREY. *The Reliability of Quarterly Marks in the Seventh Grade of Junior High School, Together with the Value of Certain Standard Tests in Predicting Them.* Johns Hopkins University Studies in Education, No. 17. Baltimore: Johns Hopkins Press, 1931. 54 p.
438. TERMAN, LEWIS M., and OTHERS. *Genetic Studies of Genius.* Vols. 1-2. Stanford University, Calif.: Stanford University Press, 1925-26.
439. TERMAN, LEWIS M. *The Intelligence of School Children.* Boston: Houghton Mifflin Co., 1919. 317 p.
440. TERMAN, LEWIS M. *The Measurement of Intelligence.* Boston: Houghton Mifflin Co., 1916. 362 p.
441. THORNDIKE, EDWARD L. *An Introduction to the Theory of Mental and Social Measurements.* 2d ed. rev. and enl. New York: Teachers College, Columbia University, 1913. 277 p.
442. THURSTONE, L. L. *The Fundamentals of Statistics.* New York: Macmillan Co., 1925. 237 p.
443. TIEGS, ERNEST W., and CRAWFORD, CLAUDE C. *Statistics for Teachers.* Boston: Houghton Mifflin Co., 1930. Chapter 2, "Labor-Saving Devices and Equipment," p. 12-27.
444. TOOPS, HERBERT A. *Ohio College Association Intelligence Tests for Entrance—Printer-scored Intelligence Tests.* Columbus: Ohio State University, 1932.
445. TRABUE, MARION R. *Measuring Results in Education.* New York: American Book Co., 1924. 492 p.

446. TYLER, RALPH W. "A Generalized Technique for Constructing Achievement Tests." *Educational Research Bulletin (Ohio State University)* 10: 199-208; April 15, 1931.
447. TYLER, RALPH W. "Measuring the Results of College Instruction." *Educational Research Bulletin (Ohio State University)* 11: 253-60; May 11, 1932.
448. TYLER, RALPH W. "What High-School Pupils Forget." *Educational Research Bulletin (Ohio State University)* 9: 490-92; November 19, 1930.
449. TYNDALL, SARA E. "How Should Committees Proceed to Attain Valid Standards for Junior-High-School Promotion?" *Junior High Clearing House* 3: 29; April-May, 1929.
450. UHL, W. L. "The Use of Standardized Materials in Arithmetic for Diagnosing Pupils' Methods of Work." *Elementary School Journal* 18: 215-18; November, 1917.
451. U. S. DEPARTMENT OF THE INTERIOR, OFFICE OF EDUCATION. *Bibliography of Research Studies in Education, 1926-1927.* Bulletin, 1928, No. 22. Washington, D. C.: Government Printing Office, 1929. 162 p.
452. U. S. DEPARTMENT OF THE INTERIOR, OFFICE OF EDUCATION. *Bibliography of Research Studies in Education, 1927-1928.* Bulletin, 1929, No. 36. Washington, D. C.: Government Printing Office, 1930. 225 p.
453. U. S. DEPARTMENT OF THE INTERIOR, OFFICE OF EDUCATION. *Bibliography of Research Studies in Education, 1928-1929.* Bulletin, 1930, No. 23. Washington, D. C.: Government Printing Office, 1930. 308 p.
454. U. S. DEPARTMENT OF THE INTERIOR, OFFICE OF EDUCATION. *Bibliography of Research Studies in Education, 1929-1930.* Bulletin, 1931, No. 13. Washington, D. C.: Government Printing Office, 1931. 475 p.
455. VAN WAGENEN, M. J. *Educational Diagnosis and the Measurement of School Achievement.* New York: Macmillan Co., 1926. 276 p.
456. WASHBURNE, CARLETON W. *Adjusting the School to the Child.* Yonkers-on-Hudson, N. Y.: World Book Co., 1932. 189 p.
457. WASHBURNE, CARLETON W. "The Grade Placement of Arithmetic Topics: a 'Committee of Seven' Investigation." *Report of the Society's Committee on Arithmetic.* Twenty-Ninth Yearbook. National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1930. p. 641-70.
458. WASHBURNE, CARLETON W., and RATHS, LOUIS EDWARD. "The High-School Achievement of Children Trained under the Individual Technique." *Elementary School Journal* 28: 214-24; November, 1927.
459. WASHBURNE, CARLETON W.; VOGEL, MABEL; and GRAY, WILLIAM S. *A Survey of the Winnetka Public Schools.* Bloomington, Ill.: Public School Publishing Co., 1926. 135 p.
460. WASHBURNE, CARLETON W., and MORPHETT, MABEL VOGEL. "Unfamiliar Situations as a Difficulty in Solving Arithmetic Problems." *Journal of Educational Research* 18: 220-24; October, 1928.
461. WATSON, GOODWIN B. "Character Tests and Their Applications through 1930." *Review of Educational Research* 2: 185-270; June, 1932. Washington, D. C.: American Educational Research Association, a department of the National Education Association.
462. WHEAT, HARRY GROVE. *The Relative Merits of Conventional and Imaginative Types of Problems in Arithmetic.* Contributions to Education, No. 359. New York: Teachers College, Columbia University, 1929. 123 p.
463. WILSON, GUY M., and HOKE, KREMER J. *How to Measure.* New York: Macmillan Co., 1928. 597 p.
464. WITMER, ELEANOR M. "Educational Research: A Bibliography on Sources Useful in Determining Research Completed or Under Way." *Teachers College Record* 33: 335-40; January, 1932.
465. WOOD, BEN D., and FREEMAN, FRANK N. *An Experimental Study of the Educational Influences of the Typewriter in the Elementary School Classroom.* New York: Macmillan Co., 1932. 214 p.
466. ZIEGLER, CAROLINE. *Test-Determined Teaching of Two Elements in Literary Appreciation.* Unpublished master's thesis, Johns Hopkins University, 1930.
467. ZIRBES, LAURA. *Comparative Studies of Current Practice in Reading.* Contributions to Education, No. 316. New York: Teachers College, Columbia University, 1928. 229 p.